COGNITION AND INSTRUCTION, *21*(1), 1–78

Learning About Statistical Covariation

Paul Cobb and Kay McClain
Vanderbilt University

Koeno Gravemeijer
Freudenthal Institute
University of Utrecht, Utrecht, The Netherlands and
Vanderbilt University

In this article, we report on a design experiment conducted in an 8th grade classroom that focused on students' analysis of bivariate data. Our immediate goal is to document both the actual learning trajectory of the classroom community and the diversity in the students' reasoning as they participated in the classroom mathematical practices that constituted this trajectory. On a broader level, we also focus on the learning of the research team by documenting the conjectures about the students' statistical learning and the means of supporting it that the research team generated, tested, and revised on-line while the experiment was in progress. In the final part of the article, we synthesize the results of this learning by proposing a revised learning trajectory that can inform design and instruction in other classrooms. In doing so, we make a contribution to the cumulative development of a domain-specific instructional theory for statistical data analysis.

Our purpose in this article is to report on a 14-week classroom design experiment conducted with a group of eighth-grade students that focused on statistical covariation. In presenting the analysis, we focus on both the trajectory of the students' learning and our own learning about the means of supporting the students' learning. This latter issue, the learning of the research team, is central to the design experiment methodology, in that conjectures about the trajectory of students' learning are tested and modified on-line while experimenting in a

Requests for reprints should be sent to Paul A. Cobb, Peabody College, Vanderbilt University, Box 330, Nashville, TN 37203. E-mail: paul.cobb@vanderbilt.edu

classroom. As we will clarify, the theoretical insights gained while testing and revising conjectures in this manner constitute the primary rationale for the instructional design that is crafted in the course of a design experiment (Gravemeijer, 1994).

In the first section of the article, we discuss the design experiment methodology by highlighting its key feature, tightly integrated cycles of instructional design and classroom analysis. We then describe the data sources and clarify the analytical method we followed. We argue that, for our purposes, a focus that encompasses both the mathematical learning of the classroom community and the reasoning of individual students as they participate in the activities of this community constitutes an appropriate unit of analysis. Against this background, we describe the setting and general organization of the classroom before outlining the hypothetical learning trajectory that we formulated when preparing for the design experiment. We next develop an analysis of the actual learning trajectory that was realized in the classroom during the design experiment. In doing so, we describe the process by which we tested and revised conjectures about the learning both of the classroom community and of individual students. In the final section of the article, we reflect back on the entire experiment in order to propose a new learning trajectory that synthesizes what we learned.

THE DESIGN EXPERIMENT METHODOLOGY

The process of conducting a classroom design experiment can be divided into three broad phases: preparing for the experiment, experimenting in the classroom, and conducting a retrospective analysis (P. Cobb, 2000; Confrey & Lachance, 2000; Simon, 2000). Gravemeijer (1994) described the preparation phase in some detail and clarified that the research team initially conducts an anticipatory thought experiment. In doing so, members of the team envision how mathematical activity and discourse might evolve as proposed types of instructional activities are enacted in the classroom, thereby developing conjectures about both (a) possible trajectories for mathematical learning and (b) the means that might be used to support and organize that learning. These conjectures compose what Simon (1995) termed a hypothetical learning trajectory, which serves to provide an orientation during the second phase of experimenting in the classroom.

In contrast to some types of classroom-based research, the primary motive for conducting a design experiment is not to assess the effectiveness of an initial instructional design formulated in advance. Instead, the intent is to improve the initial instructional design by testing and modifying conjectures about the course of the classroom community's and the participating students' mathematical learning (Brown, 1992; P. Cobb, 2001; Collins, 1999; Suter & Frechtling, 2000). To fulfill this agenda, all members of the research team[1] meet after each classroom session to discuss their interpretations of classroom events and to propose possible modifications to the learning trajectory. As a consequence, although we outline possible types of instructional activities when preparing for an experiment, the specific instructional activities used in the classroom are developed only a day or two in advance. Gravemeijer (1994) introduced the term *minicycles* to refer to these tightly integrated cycles of design and analysis. As he noted, the hypothetical learning trajectory evolves throughout a design experiment as a consequence of daily decisions even as it serves to frame and orient those local decisions.[2] It is by means of this process of ongoing adjustment and revision that the actual learning trajectory is realized in the classroom during a design experiment. Once the experiment is completed, the divergence of the actual trajectory from the trajectory hypothesized at the outset together with the justifications for the changes provide a record of the research team's learning as it enacted the daily minicycles.

The issues that arise while a design experiment is in progress are typically pragmatic and relate directly to the goal of supporting the participating students' learning. In contrast, the intent when conducting a retrospective analysis is to contribute to the development of a domain-specific instructional theory that can feed forward to guide instruction in other classrooms. A local instructional theory of this type consists of (a) a demonstrated learning route that culminates with one or more significant mathematical ideas and (b) substantiated means of supporting and organizing learning along that trajectory. As Steffe and Thompson (2000) clarified, it is this domain-specific theory that makes the results of a design experiment potentially generalizable even though they are grounded in the particulars of a single classroom. In Steffe and Thompson's terms, this is generalization by means of an explanatory framework rather than by means of a representative sample, in that insights and understandings gained when developing the retrospective analysis can inform the interpretation of events and thus pedagogical judgment in other classrooms. As we will illustrate, these insights and understandings can also feed forward to guide the formulation of a new hypothetical learning trajectory for a follow-up design experiment. Gravemeijer (1994) called these cycles of design and analysis that span an entire design experiment *macrocycles* to distinguish them from the

[1] In addition to the authors, the members of the research team for the covariation design experiment were Jose Cortina, Lynn Hodge, Maggie McGatha, Kazu Nunokawa, Nora Shuart, and Carrie Tzou. Cliff Konold and Erna Yackel served as long-term consultants and visited the classroom approximately once every 2 weeks throughout the experiment.

[2] This process of global plans evolving in response to local decisions that they serve to orient is compatible with Suchman and Trigg's (1993) analysis of design and with Lave's (1988) description of problem solving as a gap-closing process. It is also compatible with Nemirovsky and Monk's (2000) discussion of creative mathematical activity as a process of trail making, which involves moving toward a potentially revisable goal by making ongoing judgments about how to deal with immediate circumstances.

daily minicycles. Thus, although we will document the local decisions we made while the covariation design experiment was in progress, it should be clear that our intent in doing so is to contribute to the development of a local instructional theory. It is also with this in mind that we will conclude by drawing on what we learned to outline a revised learning trajectory that might provide a basis for future work. In general, this view of local instructional theories as emerging over the long term as a sequence of macrocycles is enacted is consistent with the characterization of reform as an iterative process of continual improvement (Stigler & Hiebert, 1999).

DATA SOURCES AND METHOD OF ANALYSIS

The data generated in the course of the 14-week design experiment that focused on statistical covariation included (a) video recordings made with two cameras of each of the 41 classroom sessions, (b) copies of all the students' written work, (c) two detailed sets of classroom field notes, and (d) audio recordings of all research team meetings. One of the challenges when analyzing classroom data of this type is to clarify the unit of analysis. It is, for example, tempting to characterize the shifts in activity and meaning that occur in a design experiment by speaking of changes in the students' reasoning, thereby implying that they have all reorganized their activity in the same way. However, in our view, an analytic approach of this type is potentially misleading, in that we know only too well that there are significant qualitative differences in individual students' reasoning at any point in time. To circumvent this difficulty, we take the microculture established by the classroom community as a unit of analysis while acknowledging that students participate in the communal activities that constitute this microculture in a range of diverse ways. Our goal in analyzing the actual learning trajectory of the classroom community is therefore to document the evolution of communal classroom processes that constitute the immediate social context of all the individual students' learning. We contend that when researchers and designers seem to imply that all the students have reorganized their reasoning in the same way, we can make their claims more intelligible by recasting their comments in terms of claims about the social context of all the students' learning. This reframing rejects highly questionable assertions about the homogeneity of all the students' reasoning in favor of empirically grounded claims about the conditions for the possibility of all the students' learning.[3]

[3] It should be clear in taking this approach that we are in no way ruling out the detailed analysis of individual students' reasoning. It is in fact because we acknowledge the diversity of students' reasoning that we question claims that seem to imply relative homogeneity. In the hands of a skillful teacher, this diversity can be a primary motor of the collective mathematical learning of the classroom community.

The analytical approach that we take differentiates among three distinct types of classroom norms: classroom social norms, sociomathematical norms, and normative mathematical meanings.[4] Briefly, an analysis of the social norms established by a classroom community serves to document what Erickson (1986) and Lampert (1990) termed the classroom participation structure. Examples of social norms for whole-class discussions include the obligations that students explain and justify solutions, attempt to make sense of explanations given by others, indicate understanding or nonunderstanding, and ask clarifying questions or challenge alternatives when differences in interpretations have become apparent. As these examples illustrate, classroom social norms are not specific to mathematics, but instead apply to any subject matter area. For example, one might hope that students would explain their reasoning in science or history classes as well as in mathematics. In contrast, sociomathematical norms focus on regularities in classroom actions and interactions that are specific to mathematics (Hershkowitz & Schwartz, 1999; McClain & Cobb, 2001a; Sfard, 2000a; Simon & Blume, 1996; Voigt, 1995; Yackel & Cobb, 1996). Examples of sociomathematical norms include the criteria that are established in a particular classroom for what counts as a different mathematical solution, a sophisticated mathematical solution, and an efficient mathematical solution, as well as for what counts as an acceptable mathematical explanation.

If sociomathematical norms are specific to mathematics, then normative mathematical meanings are, by definition, specific to particular mathematical ideas and are thus concerned with the emergence of what is traditionally called mathematical content. For example, in the case at hand, we will be concerned with the ways of talking and reasoning about bivariate data that became normative in the design experiment classroom. These normative meanings, it should be noted, do not correspond to an overlap in the teacher's and students' individual interpretations (Voigt, 1985). Any attempt to delineate an overlap of this type takes individuals as the unit of analysis in that the focus is on the relation among individual interpretations. In contrast, inferences about normative interpretations take the classroom community as the unit of analysis and attempt to delineate meanings that are constituted as legitimate in the classroom. Such meanings are communal rather than individual accomplishments in that their status of legitimacy is established collectively by the teacher and students.

It is important to emphasize that when we analyze classroom video recordings, we cannot see the classroom community as a discrete, concrete entity in the same

[4] For the purposes of this article, we have simplified our interpretive framework by speaking of normative mathematical meanings rather than classroom mathematical practices. As we described and illustrated elsewhere (P. Cobb, Stephan, McClain, & Gravemeijer, 2001), a classroom mathematical practice comprises three interrelated types of mathematical norms: a normative purpose for engaging in mathematical activity, normative standards of argumentation, and normative ways of reasoning with tools and symbols.

way that we can see the teacher and students as distinct physical beings. As a consequence, we cannot observe normative mathematical meanings directly any more than we can directly observe the meanings that an individual student's data analysis activity has for her or him. We described and illustrated the methodological approach we take when inferring normative meanings in some detail elsewhere (P. Cobb et al., 2001). For our purposes, it suffices to note that normative ways of reasoning and acting are not mere arbitrary conventions for members of a community that can be modified at will (Sfard, 2000a). Instead, these are ways of reasoning and acting that are constituted as legitimate or acceptable within a community.[5] Consequently, in analyzing the data generated during the covariation design experiment, we developed and tested conjectures about normative ways of talking and reasoning about data by focusing on the status that students' contributions came to have in the classroom. For example, we documented whether the students were obliged to give a justification when they first organized data in a novel way. The need to give a justification is one indication that this particular way of structuring data was still open to question. Evidence that students who later organized data in this way no longer needed to give a justification indicated that this way of reasoning with data might have become normative. Beyond this, we opened our inferences to the possibility of refutation by searching for instances where a student appeared to have violated a way of reasoning that we conjectured was normative (Much & Schweder, 1978). In each of these instances, we focused on the status that the student's contribution came to have as the classroom discourse progressed. We, of course, had to revise our conjecture in those cases where the student's contribution was constituted as legitimate.

Thus far, our discussion of methodological issues has focused on the classroom microculture as a unit of analysis. A second challenge that arises when analyzing a large, longitudinal dataset of the type generated during a design experiment is to develop a way of working through the data systematically so that the resulting account is credible. In this regard, it is worth noting that analyses that locate students' mathematical activity in social context often deal with only a few lessons, or perhaps focus on just a few minutes within one lesson. Detailed analyses of this type can clearly make an important contribution to design research. However, a methodological issue that we have sought to address is that of developing a way to account for the collective mathematical learning of the classroom community not during a 10-min episode, but over the entire time period spanned by a design experiment. Consequently, although the account we will offer of the covariation design experiment is grounded in

[5] This observation also bears directly on the process by which the members of a community develop distinct identities as they participate in the continual regeneration of communal norms (Schutz, 1962; Wenger, 1998). In particular, members of a community do not merely act in accord with the norms of a community. They become people who consider that their rights have been infringed when they perceive that a norm has been breached.

the details of specific classroom events, it necessarily lacks the immediacy of more closely circumscribed analyses. We argue that this trade-off is justifiable, given that the process of supporting the development of significant mathematical ideas typically takes a period of weeks or months.

We follow the standard convention for reporting interpretivist analyses by presenting a limited number of representative episodes to clarify the primary assertions that emerged while conducting the analysis (Atkinson, Delamont, & Hammersley, 1988; Taylor & Bogdan, 1984). It is therefore important to emphasize that these assertions about the classroom community's mathematical learning did not typically arise from the analysis of a single episode. Instead, as we illustrated elsewhere (P. Cobb et al., 2001), the interpretation of specific episodes and the delineation of general assertions are interdependent in that each informs the other. As a consequence, the interpretations we will propose of particular episodes are located within a network of mutually reinforcing inferences that span the entire dataset.

We have already indicated that the analytical method we take treats inferences about normative ways of reasoning as conjectures that are open to refutation. This approach is a variant of Glaser and Strauss's (1967) constant comparison method as adapted to the needs of design research (P. Cobb & Whitenack, 1996). Glaser and Strauss's (1967) method treats data as text and aims to develop coherent, trustworthy analyses of their possible meanings. The hallmark of their method is that as new classroom episodes are analyzed, they are compared with currently conjectured themes or categories. This process of constantly comparing episodes leads to the ongoing refinement of the theoretical categories that remain grounded in the data. As Glaser and Strauss noted, negative cases that appear to contradict a current category are of particular interest and are used to further refine the emerging categories.

The specific analytical approach that we followed has two main phases (P. Cobb et al., 2001). In the first phase, we worked through the data generated during the covariation design experiment chronologically, episode by episode, where the determining characteristic of an episode was that a single mathematical theme was the focus of the teacher's and students' activity and discourse. In doing so, we developed conjectures about ways of reasoning and communicating that might have been normative in the classroom at a particular point in time. The result of this first phase was a chain of conjectures, refutations, and revisions that was grounded in the details of specific episodes. In the second phase, the record of the first phase itself became data that were (meta-) analyzed to develop a reasonably succinct, empirically grounded chronology of the mathematical learning of the classroom community. During this phase, we scrutinized the conjectures developed during the first phase about the possible emergence of normative meanings from a relatively global perspective that looked across the entire design experiment. The resulting analysis of the evolution of normative meanings over the course of the design experiment then provides an account of the actual learning trajectory of the classroom community.

THE SETTING OF THE DESIGN EXPERIMENT

The design experiment was conducted over a 14-week period in an urban middle school in the fall of 1998 and involved 41 classroom sessions of approximately 40 min duration. During the fall of the previous year, we conducted a design experiment in a seventh-grade classroom in the same school that focused on the analysis of univariate data. Our intent was to work with the same group of 29 students during the first part of their eighth-grade year to investigate the analysis of bivariate data with a particular emphasis on statistical covariation. However, the teacher of the eighth-grade mathematics classes felt that she could not justify the time away from instruction. Her students were required to take an end-of-grade exam and she felt pressure to address concepts from that test. Her curriculum was relatively prescriptive in that it specified the mathematical skills and concepts that should be taught throughout the school year. We were therefore unable to conduct the eighth-grade design experiment during regular mathematics periods and asked for student volunteers to work with us during their afternoon activity period. Of the original 29 students, 8 had transferred to other schools and 4 had other obligations (e.g., practice for the school play or for the school band). Of the remaining 17 students, 16 volunteered to give up their activity period and 11 continued to attend throughout the 14 weeks of the experiment. The 5 students who dropped out, all of whom were White, indicated that they were having difficulty completing their homework for other classes and wanted to use the activity period for this purpose. Seven of the 11 students who participated for the entire experiment were African American, 3 were White, and 1 was Asian American. An analysis of interviews conducted with all of the original 29 students at the end of seventh-grade design experiment indicated that these 11 students were reasonably representative of the entire group in terms of the ways in which they reasoned about data.

CLASSROOM ORGANIZATION

During the design experiment, Kay McClain assumed the primary teaching responsibilities and was assisted on occasion by Paul Cobb. For ease of explication, we will not differentiate between their contributions to classroom discourse, but will instead speak simply of the teacher. The flow of classroom activities typically had the following structure, which might span two or more class sessions: (a) a whole-class discussion in which the teacher and students talked through the data creation process, (b) individual or small-group activity in which the students worked at computers to analyze data, and (c) a whole-class discussion of the students' analyses. The rationale for the first phase of this activity structure stemmed from our prior work with the students during their seventh-grade year (McGatha, Cobb, & McClain, 1999). In the prior design experiment, it proved crucial that datasets had a history for the students such that they were grounded in the situation from which they were generated and that they reflected the particular interests and purposes that had led to their creation (Latour, 1987; Lehrer & Romberg, 1996; Roth, 1997). It was to this end that the teacher talked through the process of generating the data with the students. This involved discussing the particular problem or question to be investigated, clarifying its social or scientific significance, delineating aspects of the situation that might be relevant to the question at hand, and developing procedures for measuring them. The data the students were to analyze were then introduced as having been generated in this way. An analysis of students' participation in the data creation process across the two design experiments indicates that the approach we took was generally successful (Tzou, 2000). There was a clear handover of responsibility from the teacher to the students in the course of which the students initiated discussions with increasing frequency about the need to control extraneous variables and about sampling methods.

Once the teacher and students had resolved issues about the process of generating the data, the students analyzed the data and wrote a short report for a person who would make a policy decision based on their analysis. In the first eight classroom sessions, the students used a computer minitool from the seventh-grade experiment with which they were already familiar to analyze univariate datasets. In the next five classroom sessions, the teacher and students developed ways of inscribing bivariate data before a new computer minitool for analyzing bivariate data was introduced in the 14th classroom session. This minitool was used during most of the remaining sessions of the experiment. The students conducted their analyses either individually or in pairs as they chose, subject to the constraint that the class had access to a total of eight computers.

The research team prepared for the final phase of the classroom activity structure, the whole-class discussions of the students' analyses, by developing conjectures about mathematically significant issues that might, with the teacher's proactive guidance, emerge as topics of conversation. Our intent was to capitalize on the diversity in the students' reasoning about data by identifying analyses that, when discussed directly or compared with other analyses, might lead to substantive mathematical conversations that advanced our pedagogical agenda. As a consequence, whole-class discussions were not viewed merely as opportunities for students to share their reasoning. Instead, they focused on selected ways of reasoning that could be justified in terms of their contribution to the realization of a potentially revisable learning trajectory. A computer projection system that enabled the students to demonstrate how they had structured particular datasets was used to support these discussions throughout the design experiment.

THE HYPOTHETICAL LEARNING TRAJECTORY

The process of formulating a hypothetical learning trajectory that provides an initial orientation for a design experiment involves specifying (a) the significant mathematical ideas that constitute the potential developmental endpoints, (b) the anticipated starting points, and (c) the envisioned learning route and means of support. We discuss each of these three aspects of the hypothetical trajectory in turn for the covariation design experiment.

Potential Endpoints

Our overall goal in the design experiment was to support the emergence of increasingly sophisticated ways of analyzing bivariate data as part of the process of developing effective data-based arguments. The overarching mathematical idea that served to orient our design effort was that of bivariate distribution. We wanted the students to come to view bivariate datasets as distributed within a two-dimensional space of values (Wilensky, 1997). Notions such as the direction and strength of the relationship between the two sets of measures generated when creating the data would then emerge as ways of describing how the data are distributed within this space of values.

Our reference to a two-dimensional space of values indicates the central role that we attributed to the scatter plots as a way of inscribing bivariate data. The image of classroom discourse that we had in mind was that of scatter plots coming to be talked about and referred to as texts of situations from which the data were generated. We therefore wanted it to become normative that aspects of the situation that were measured when generating the data covary in some way, and that the nature of that covariation can be read from a scatter plot. In considering what might be involved in reasoning about a scatter plot in this way, we took account of the observation that students frequently read graphs of this type diagonally rather than vertically, focusing on the distance of points from a line of best fit rather then on deviations in the *y* direction (Clifford Konold, personal communication, July 18, 1998). For these students, it is the line of best fit rather than the two sets of measures that have been plotted orthogonally that constitutes the frame of reference. We conjectured that, in contrast, proficient data analysts view the graph as organized into vertical slices, each of which can be viewed as the (univariate) distribution of the measures of one quantity for an interval of values of the other quantity. In the example shown in Figure 1, the process of reading the graph involves discerning trends and patterns in the distribution of average SAT scores as measures of expenditure increase.[6]

[6] Support for this view comes from a study conducted by Noss, Pozzi, and Hoyles (1999). They reported that a group of nurses with whom they conducted a short design experiment initially had difficulty in seeing any relationships in scattergrams (e.g., of systolic blood pressure and age). However, when the nurses partitioned the data into vertical slices corresponding to age groups and found group means, they were able to discern relationships. Further, some of the nurses went on to explore box plots of the grouped data on their own initiative as a means of "taming" the distribution of the data.

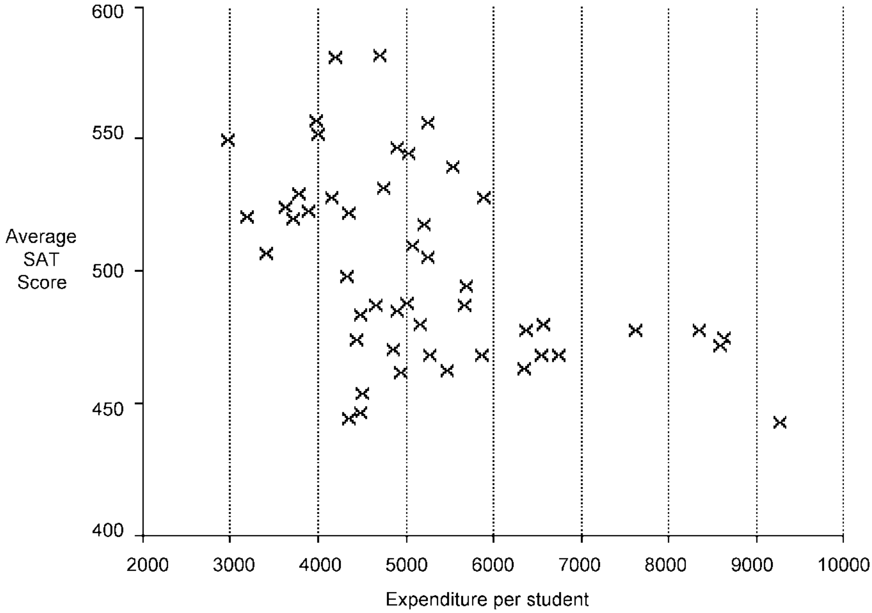


FIGURE 1 Average SAT scores and expenditures per student for the United States.

Our claim is not, of course, that skilled readers of scatter plots consciously partition graphs into slices. Instead, we conjectured that the perceptual activity of skilled readers implicitly involves tracking the distribution of measures of the *y* quantity as they scan across the graph. This process of scanning across a scatter plot vertically rather than diagonally is explicit in procedures for finding the line of best fit (i.e., minimize the squares of the deviations of the *y* measures). As will become apparent, this view of a bivariate distribution as a distribution of univariate distributions strongly influenced the design of the new computer minitool that we introduced during the design experiment.

Starting Points

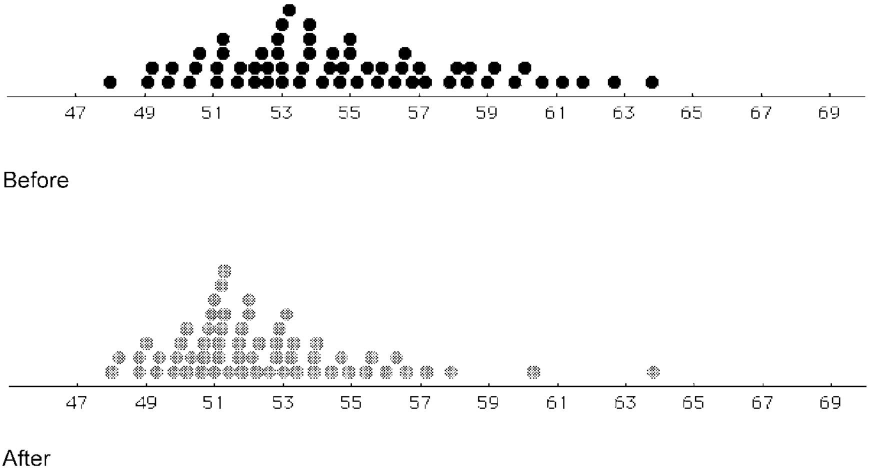
To clarify the starting points for the covariation design experiment, we refer to the prior design experiment that we conducted when the students were seventh graders. As we noted, the instructional activities used in this experiment involved analyzing univariate datasets. The means of supporting the students' learning included two computer minitools that we developed when preparing for the experiment. The second of these tools, which was used in the latter part of the design experiment, enabled students to analyze one or two univariate datasets of up to 400 data points. Individual data points were inscribed as dots on a horizontal axis of values (see Figure 2).

The tool provided the students with a variety of options for structuring datasets. The first, called “Create Your Own Groups,” involved dragging vertical bars along the axis to partition the datasets into groups of points. The remaining four options were:

- Partitioning the data into groups of a Specified Size (e.g., 10 data points in each group).
- Partitioning the data into groups with an Equal Interval Width (i.e., a precursor to histograms).
- Partitioning the data into Two Equal Groups.
- Partitioning the data into Four Equal Groups (i.e., a precursor of box plots).

In each of these options, the students could elect to hide the individual data points or leave them visible. If the Hide Data option is chosen, the dots signifying individual data points disappear, leaving only vertical partition lines.

Analyses of the classroom community’s actual learning trajectory in this design experiment can be found in P. Cobb (1999), McClain and Cobb (2001b), and McClain, Cobb, and Gravemeijer (2000). Briefly, the metaphor of a hill to describe the shape of datasets such as those shown in Figure 2 emerged shortly after the students began using the second minitool. In addition, the interpretation that



Before

After

FIGURE 2 Data inscribed as line plots in the second minitool.

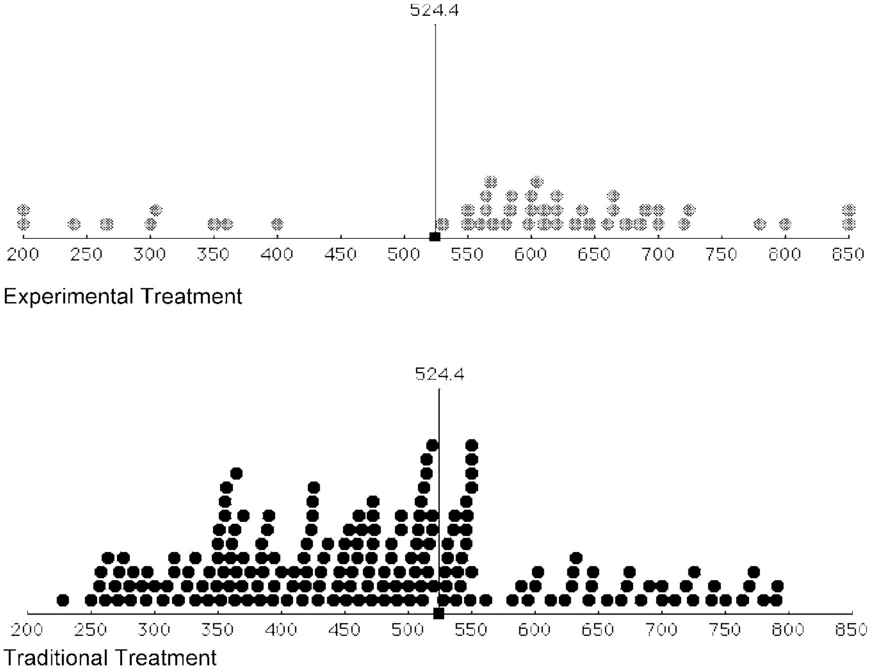


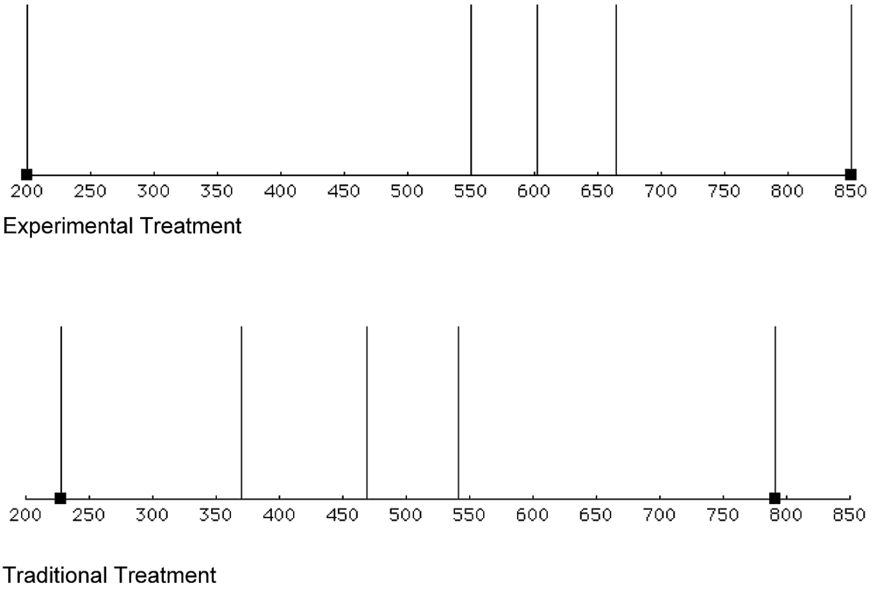
FIGURE 3 The AIDS protocol data partitioned at T-cell counts of 525.

the data in a particular interval were a qualitative proportion of the entire dataset rather than a mere additive part of it had become normative by the end of the experiment. As an example, Figure 3 shows the T-cell counts for 46 patients who enrolled in an experimental protocol for the treatment of AIDS and 186 patients who enrolled in a standard treatment protocol.[7] Both datasets have been partitioned at the T-cell count of 525 by using the Create Your Own Groups option. In a case such as this, the students routinely spoke of the "majority" of the data or "most of the people" being above a T-cell count of 525 in the experimental treatment and below this value in the standard treatment. In these conversations, there was every indication that the meaning of terms such as *the majority* and *most* as signifying a qualitative proportion of a dataset and thus a qualitative relative frequency was normative.

In the example shown in Figure 3, the students who partitioned the datasets at T-cell counts of 525 did so because what they referred to as the "hill" in the experimental protocol data was above 525, whereas the hill in the standard

[7] These datasets were introduced only after a lengthy discussion of the data creation process.

treatment data was below 525. In reasoning in this way, they used the second minitool to identify and describe perceptually based patterns in the data. In contrast to this perceptually based reasoning, a number of the students compared the two treatment programs by using the minitool to organize datasets independently of visual features. Consider, for example, Figure 4, in which the AIDS data are partitioned into Four Equal Groups and the individual data points are hidden. In this case, the partitions do not isolate perceived clumps or hills in the data that could be read as qualitative proportions. Instead, the graphs serve as a means of comparing and contrasting the ways in which the two sets of data are distributed. A series of discussions conducted near the end of the seventh-grade design experiment indicated that reasoning of this type in which the distribution of data is inferred from a graph was yet to become normative. It was also noticeable that, in these discussions, the teacher and students rarely spoke of hills or used other terms that referred to the shape of datasets when they reasoned about graphs of Four Equal Groups. Instead, they typically talked about these graphs of Four Equal Groups by referring to the percentage of the data above or below a particular value or within a particular interval. In the case of Figure 4, for example, some of the students noted that the T-cell counts of the lowest 75% of the patients in the standard protocol were in approximately the same interval as the counts of only the lowest 25% of the patients in the experimental protocol (i.e., the ranges of the lowest quartile



Experimental Treatment

Traditional Treatment

FIGURE 4 The AIDS protocol data organized into Four Equal Groups with data hidden.

of the experimental treatment and the lowest three quartiles of the standard treatment were approximately the same). This way of speaking and reasoning about data organized into Four Equal Groups might well have reflected the fact that the datasets the students analyzed in the latter part of the seventh-grade experiment were typically not particularly smooth and thus did not lend themselves to being described in terms of shape. As will become apparent, this exclusion of descriptions cast in terms of shape had unanticipated consequences in the eighth-grade design experiment.

In summary, we concluded from our analyses of the seventh-grade design experiment that the basis for communication at the beginning of the eighth-grade experiment might include:

- Using the hill metaphor to describe the shapes of datasets inscribed as line plots.
- Comparing univariate datasets by structuring them in terms of perceptually based patterns.
- Reasoning multiplicatively about datasets structured in this manner in terms of qualitative proportions.

In addition, we anticipated that an appreciable number of the students would readily partition data into Equal Interval Widths or Four Equal Groups to compare how two sets of data were distributed. Interviews conducted shortly after the seventh-grade experiment was completed indicated that 19 of the 29 students could use graphs of Equal Interval Widths and of Four Equal Groups in which the data were hidden to develop effective data-based arguments (P. Cobb, 1999). Eight of these 19 students were among the 11 who participated in the entire eighth-grade experiment.

The Conjectured Learning Route and Means of Support

We stress that because there was a 9-month gap between the two experiments, we viewed our assumptions about the starting points for the eighth-grade experiment as conjectures. The initial activities we planned were therefore designed to serve as performance assessment tasks and involved using the second computer minitool to compare two univariate datasets. If necessary, we intended to revisit issues that had been the focus of discussions in the latter part of the seventh-grade experiment before moving to the analysis of bivariate data.

Our immediate goal when the students began to analyze bivariate data was to support the development of ways of inscribing the data. At a minimum, we wanted to ensure that the students viewed the inscriptional form of scatter plots as a solution to a problem that they considered significant. To this end, we planned to ask the students to develop a graph or a diagram of a bivariate dataset that would enable them to make a decision or a judgment. Although we expected that some of the students might create inscriptions similar to scatter plots by drawing orthogonal axes, we also anticipated that some might develop double-bar graphs or other potentially less useful inscriptional forms. To frame the discussion of the students' graphs, we planned to raise the issue of the extent to which their various inscriptions enabled them to assess how one of the measured quantities varied as the other increased. It was only when the relevance of this criterion had become normative that we planned, if necessary, to introduce the scatter plot as a way of inscribing data that made it easier to address the question at hand.

A second related issue that we considered when preparing for the design experiment concerned the importance of it becoming normative that bivariate data consist of the measures of two attributes of each of a number of cases. To address this concern, we discussed at length the type of language that the teacher might support when talking through the data creation process with the students. As we noted, this process would involve first discussing the particular question or issue to be investigated and clarifying its social or scientific significance. Against this background, the teacher would guide the delineation of aspects of the situation that were relevant and should be measured to address the issue. We conjectured that both here and in the subsequent discussions of the students' analyses, it would be important to develop ways of talking that referred explicitly to cases whose attributes had been measured rather than to speak solely in terms of the measures. This, we reasoned, might support the view that each dot on a scatter plot signifies a single case whose measures are indicated by its location with respect to the axes.

The issues we have discussed thus far relate to what we anticipated would be the introductory phase of the design experiment. It was against this background that we planned to introduce the third computer minitool, in which bivariate data are inscribed as a scatter plot. The students could adjust the scales of the axes by changing the maximum and the minimum values. We also included a feature called Dots, by which, if the students clicked on any data point, perpendiculars from the axes to the dot would be shown (see Figure 5). We anticipated that the use of this feature in whole-class discussions would aid the teacher in ensuring that discourse was about relationships between the two measures of each of a number of cases rather than about a mere configuration of dots scattered between two axes.

Beyond this simple feature, the minitool offered four differing ways of organizing bivariate data:

Cross. This option divides the data display into four cells and shows the number of data points in each cell (see Figure 6). The students could drag the center of the Cross to any location on the display, thereby changing the size of the

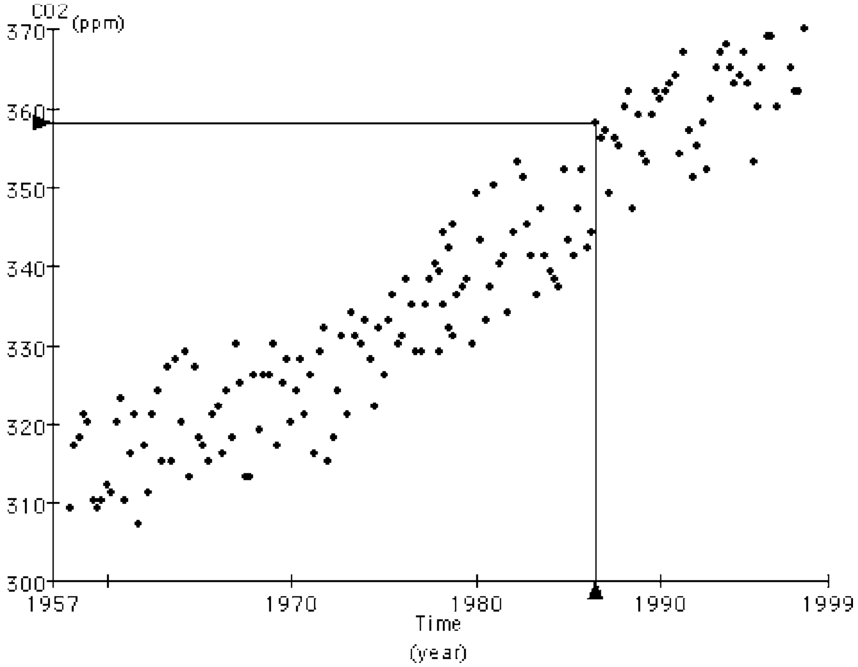


FIGURE 5 The third minitool with the Dots option activated for one data point.

cells. As they did so, the record of the number of data points in each cell adjusted automatically. In terms of the students' prior instructional history, the Cross can be viewed as the two-dimensional correlate of the Create Your Own Groups option included in the second minitool. As we noted, the students typically used this latter option to identify and describe perceptually based patterns in univariate datasets.

Grids. The students could select from a pull-down menu of Grids that ranged in size from 4 × 4 to 10 × 10. The selected Grid was shown superimposed on the data display and the number of data points in each cell was shown. The Grids option can be viewed as the two-dimensional correlate of the Equal Interval Width option included in the second minitool.

Two Equal Groups. This option partitions the data display into columns or vertical slices, the widths of which divide the horizontal axis into equal intervals (see Figure 7). The minimum number of slices that the students could choose was 4 and the maximum was 10. Within each slice, the data points are partitioned into two equal groups (i.e., the display shows the median and the low and high values

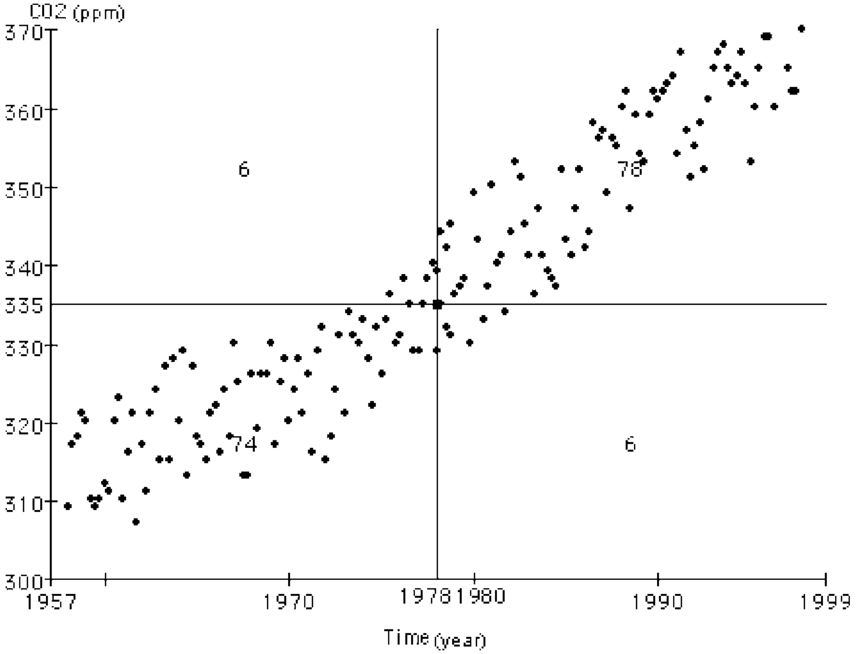


FIGURE 6 The Cross option of the third minitool.

within each slice). This option can be viewed as the two-dimensional correlate of the Two Equal Groups option included in the second minitool.

Four Equal Groups. This option is similar to the Two Equal Groups option except that the data points within each slice are partitioned into Four Equal Groups (see Figure 8). It can be viewed as the two-dimensional correlate of the Four Equal Groups option included in the second minitool.

The only remaining feature of the minitool to note is that the individual data points could be hidden. This option was designed to support conversations in which trends and patterns in the distribution of data are inferred from graphs. The remarks we made about potential relationships between the options of this tool and of the second minitool hint at our underlying rationale. At a deeper level, this rationale for the design of the third minitool stems from our conjecture that it might be productive for pedagogical purposes to view a bivariate distribution as a distribution of univariate distributions. In the case of the Grids option, for example, the cell values in each vertical slice might be read as signifying a univariate distribution. Similar comments can be made about the inscriptions within each slice of the Two Equal Groups and Four Equal Groups options. For its part, the

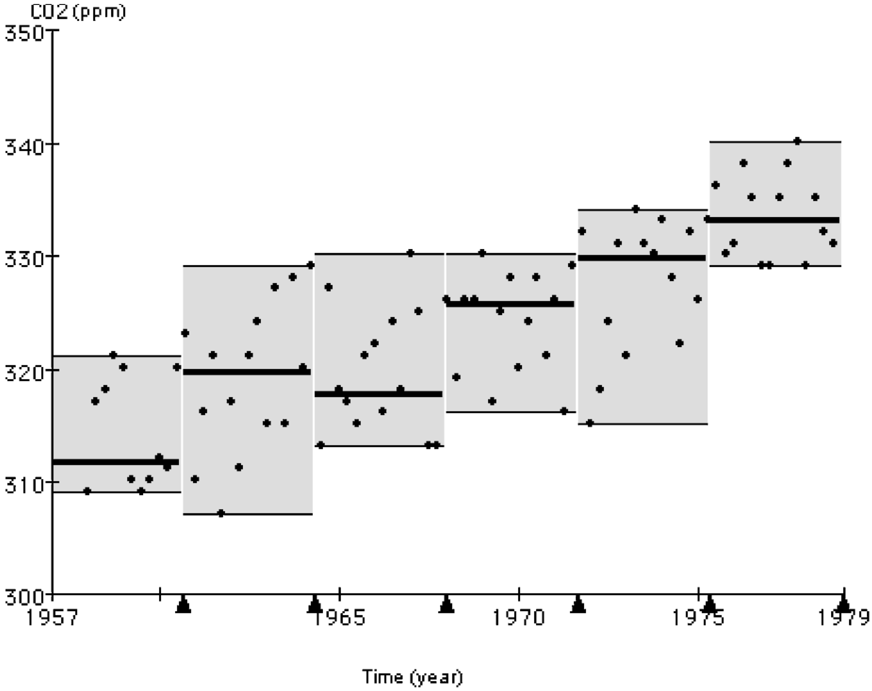


FIGURE 7 The Two Equal Groups option of the third minitool.

Cross option can be viewed as both a precursor to 2 × 2 contingency tables and a simple case of the Grids option with the added flexibility that students can vary the interval widths.

We anticipated that the envisioned learning trajectory would give rise to several challenges for the teacher. The most pressing of these involved supporting discussions in which it became normative to interpret the vertical slices in the various options as the distribution of the measures of one quantity for cases whose measures of the second quantity were within the indicated interval. Only then could the issue of how the distribution of the measures of one quantity changed as measures of the second quantity varied become a topic of conversation. We again stress that we viewed the envisioned trajectory as potentially feasible only because the teacher might be able to build on the starting points that we outlined. As we noted, these provisional starting points were themselves a consequence of the students' participation in the prior seventh-grade design experiment.

A final issue that we took account of when preparing for the design experiment concerned lines of best fit. We intended to approach this issue informally and wanted it to become normative that a line fitted through the configuration of dots

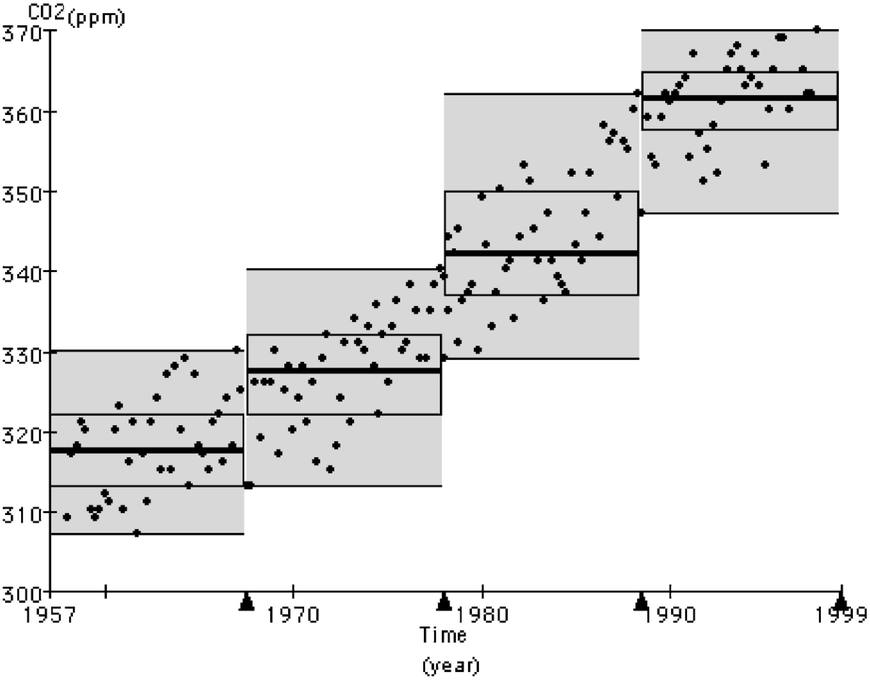


FIGURE 8 The Four Equal Groups option of the third minitool.

on a scatter plot signifies a conjectured relationship of covariation about which the data points are distributed. We therefore planned to delay line fitting until the reading of scatter plots as bivariate distributions (i.e., as distributions of univariate distributions) had become normative. As a consequence, we decided not to activate a curve-fitting option on the minitool when the students first used it, but instead to focus classroom activity on various ways of structuring and organizing bivariate datasets.

It should be clear in this discussion of the hypothetical trajectory that we viewed the development of increasingly sophisticated ways of reasoning about data as being inextricably bound up with the development and use of increasingly sophisticated ways of inscribing data (Biehler, 1993; de Lange, van Reeuwijk, Burrill, & Romberg, 1993; Lehrer & Romberg, 1996; Roth & McGinn, 1998). More generally, the approach we planned to take is broadly consistent with theoretical perspectives that treat tools and symbols as reorganizers rather than mere amplifiers of activity (Dörfler, 1993; Kaput, 1994; Meira, 1998; Pea, 1993). It should also be apparent that, in line with our discussion of the design experiment methodology, the conjectures we developed when formulating the hypothetical learning trajectory deal with the learning of the classroom community rather than of any particular students. In reporting the actual learning trajectory enacted in the course of the design experiment, we continue to focus on communal learning while also attending to the quality of individual student's reasoning.

THE ACTUAL LEARNING TRAJECTORY

In presenting the analysis, we first document the social and sociomathematical norms that were established in the design experiment classroom. Against this background, we then give an account of the actual learning trajectory by dividing it into five broad phases, each of which involves the emergence of a distinct nexus of normative meanings.[8] In discussing the first of these phases, we present sample episodes to illustrate the types of issues that typically came to the fore during data creation discussions. In discussing the remaining four phases, we present sample episodes to clarify the claims we make about the development of normative mathematical meanings. Throughout the analysis, we also attend to our own activity as well as that of the teacher and students by documenting our ongoing pedagogical decision making. This record of the process of testing and revising pedagogical conjectures will serve as a basis for our subsequent reflections and thus for the formulation of a new learning trajectory that synthesizes what we learned.

Social and Sociomathematical Norms

The social norms and sociomathematical norms that we previously documented when analyzing the seventh-grade design experiment (P. Cobb, 1999; McClain & Cobb, 2001b; McClain et al., 2000) were quickly reestablished at the beginning of the eighth-grade experiment. The social norms for whole-class discussions included the obligations that the students explain and justify their reasoning, ask clarifying questions to understand other students' reasoning, and indicate agreement or disagreement with others' arguments. These norms, which were inferred by analyzing classroom discourse, proved to be highly consistent with students' understanding of their obligations as revealed in a series of interviews conducted with them outside the classroom while the eighth-grade experiment was in progress (Hodge, 2001). These interviews were conducted as part of a separate investigation

[8] In terms we used elsewhere, each of these phases corresponds to the emergence of a distinct classroom mathematical practice (Bowers, Cobb, & McClain, 1999). The methodology for conducting an analysis of this type was discussed by P. Cobb et al. (2001).

by a member of the research team who was not directly involved in the design experiment. The following comments are representative

Interviewer: How would I be a good student [in the statistics class]?

Kim: You just ask questions about what you don't understand and tell them what you think.

Ben: Your job is to know how to express your opinion and know how to do it. And not worry about it if someone disagrees with you.

Janice: You ask questions and contribute to what we're talking about.

Sinae: You have to do a good job explaining how you looked at the problem. That's important since you didn't talk with everybody else when you were looking at the graph.

Interviewer: If I were a new student [in statistics], tell me some advice about doing well.

Martha: Talk a lot and

Interviewer: (Interrupts) About anything?

Martha: Not just anything. You talk about your way, or you add something to someone else's way. You can't just say that you agree or you disagree. Ms. M [statistics teacher] makes you explain it. You have to ask questions about things that you don't understand.

Interviewer: What do you mean?

Martha: If you, um, don't understand why someone did something you have to ask them about it. You can't just say, oh yeah, that's okay, what you did.

Interviewer: So to be a good student you have to listen and bring up good arguments?

Brad: Yeah. You have to listen and ask questions about other people's ways. That's really what you have to do. When you explain what you did you have to make sense. You can't just talk about what you ended up with.

Suzanne: You can't just talk about your conclusion because that doesn't let anybody know why you did things.

Interviewer: Is that important?

Suzanne: If you don't talk about what you were thinking about then we don't know if it all is okay…we can't figure out if it is a good point.

Clearly, the teacher's success in negotiating these obligations with the students facilitated the task of inferring the emergence of normative mathematical meanings.

For the purposes of this analysis, two of the most important sociomathematical norms are those of what counts as a different solution and as an acceptable argument. In line with our prior analysis of the seventh-grade experiment, the norm of what counted as a different solution centered on the way in which the data had been structured and interpreted rather than on the final conclusion reached during the analysis. Thus, analyses were constituted as different even if the computer minitool was used in the same way provided that the data were interpreted differently. With regard to the norm of what counted as an acceptable argument, it was not necessarily sufficient for students to explain how they had structured the data during an analysis. Instead, the students were obliged to explain their reasons for structuring the data in a particular way when this was not clear to other students. Typically, in stating these reasons, the student had to explain how a particular way of structuring the data was relevant with regard to the question or issue being addressed by the analysis. As a consequence, the classroom discourse was, for the most part, conceptual rather than calculational in nature in that students had to explain not only the process by which they arrived at a result, but the reasons for following that particular process (P. Cobb et al., 2001; Thompson, Philipp, Thompson, & Boyd, 1994).

Comparing Univariate Datasets

Turning now to consider the first phase of the actual learning trajectory, the students used the second minitool to compare univariate datasets with unequal numbers of data points in the first 8 of the 41 sessions of the design experiment. In one instructional activity, for example, the students compared the response times of two commercial ambulance companies. In talking through the data creation process, the teacher explained that the school district had to decide which of two companies would receive a contract to provide service to the school district. The first part of the data creation discussion focused on clarifying why this question was significant:

Teacher: Do you know how expensive it is to call an ambulance?

Kim: $75.

Teacher: You would say $75?

Kim: Yeah.

Teacher: The ones that I've talked to charge $125 just to show up and that's if they don't do anything.

Wes: For an ambulance?

Teacher: Yes, if you were able to crawl in to the ambulance and go to the hospital by yourself that's $125. That's a real expensive taxi ride.

Janice: So, what you're saying is that they charge it to the school?

Teacher: Yeah, yeah, just to take the students or the teacher or whoever to the hospital.

Wes: Does that count for everyone?

Teacher: Yeah, one person.

Mark: (Inaudible).

Teacher: This is just for Nashville. Ambulance service in Nashville.

Suzanne: So, basically you're paying to get your life saved?

Teacher: Yes, exactly. That's a good point.

Martha: Do they charge if you call and you don't need them?

Teacher: Yes, if they come it's $125. Just for showing up.

Martha: Even if you don't need them when they come?

Teacher: When they get there and you say false alarm. 125 bucks. If they do something. They have drivers with some medical training.

Student: Paramedics.

Teacher: Yes, paramedics. If they do something they add those charges on to the $125.

The students' need to understand the situation from which the data were generated had come to typify data creation discussions by the end of the seventh-grade experiment. As we documented elsewhere (Tzou, 2000), the students refused to begin an analysis if the phenomenon under investigation did not make sense to them.

Once the situation had been clarified to the students' satisfaction, the teacher asked them which aspects of the situation they thought should be considered when selecting one of the two companies. During this exchange, the teacher recorded the suggestions on a whiteboard:

Teacher: So, in making this decision, what kinds of things would you want to know about the ambulance company? Ben?

Ben: How fast it can get there.

Teacher: Okay, I'm going to call that response time. Is that what you were talking about?

Ben: (Nods).

Suzanne: How efficient they are.

Teacher: What do you mean by efficiency, Suzanne?

Suzanne: Like they actually know what they're doing and they don't do something that doesn't need to be done.

Teacher: So, how well trained their medical people are. Martha?

Martha: Locations to all the schools.

Teacher: Location.

Martha: Yeah, because you don't want, you want something in the middle of Nashville. Not off to the side. Like if something happens in East Nashville and you're in West Nashville.

Teacher: That's a very good point, and in fact what happens with most ambulance companies is they have what are called dispatch locations where they would have ambulances positioned all over the city.

The manner in which the students generated the resulting list of relevant issues is again typical of the data creation discussions conducted at the end of the seventh-grade experiment (Tzou, 2000). This contrasts sharply with the data creation discussions conducted at the beginning of the seventh-grade experiment in that the students had frequently told personal narratives that related to the topic at hand (e.g., accounts of incidents with which they were familiar in which an ambulance had been called to assist a friend or relative).

As was typical in data creation discussions, the teacher next explained that the people making the decision had narrowed the question down to one of the aspects that the students had identified, in this case response time.

Teacher: What do you think we mean when we say response time. Brad?
Brad: How fast, how fast they get there.
Teacher: Exactly. Is everybody okay with that? So, how would they measure response time? If they got some information on it, what would they do to find out which has got the better response?
Kim: They'd do drills.
Teacher: What do you mean by drills?
Kim: They could call the ambulance there, not on purpose. Tell them what they were doing and see how fast they get there.
Teacher: When you say fast, what would they be timing?
Kim: The speed.
Teacher: Okay, the speed. Let's hear from somebody else. Mark?
Mark: I was going to ask a question.
Teacher: Can you hold it for a second? Suzanne?
Suzanne: Well, like let's say there's ambulance two miles away from here, but it takes them a half an hour to get here. But then the other company has an ambulance that's two miles away and it takes them 5 minutes to get here.
Teacher: So you think the number of minutes is more important?
Suzanne: How long it takes compared to where they are.
Teacher: So, from when you make the phone call to when they actually get here. Is that what you were going to say, Martha?
Martha: From when they make the phone call.
Teacher: From the time they place the call to the time the ambulance arrives is called response time.

The issue addressed in the remainder of the discussion, which continued for several minutes, was that of how to generate data on the response times of the two companies. During this part of the discussion, the teacher used the computer projection system to show the students line plots of 205 response times for one company and 162 for the other (see Figure 9). Once she had done so, the students raised a number of concerns about the sampling procedures that had been used. These included the lengths of the emergency runs made by each company, the traffic conditions, and the relevance of the data to the question at hand (i.e., the relation between the location of the ambulance stations and the calls they responded to on the one hand, and the locations of schools within the city on the other). It was only when these issues had been resolved that the students moved to the computers to analyze the data.

We have discussed the process of talking through how the data might be generated for the ambulance response time activity at some length because it is representative of the data creation discussions conducted during the remainder of the eighth-grade experiment (Tzou, 2000). As the sample episodes indicate, there were no signs of regression when compared with the ways in which the students had participated in these discussions during the latter part of the seventh-grade design experiment. The concerns that the students raised about sampling procedures and the control of extraneous variables indicate that most anticipated that the conclusions that could be drawn from the data depended on the soundness of the data creation process. It is also worth mentioning that, from the students' point of view, it was crucial that both the purposes for analyzing the data and the audience for their reports be clarified. In both the latter part of the seventh-grade experiment and in the eighth-grade experiment, they considered that a norm had been violated

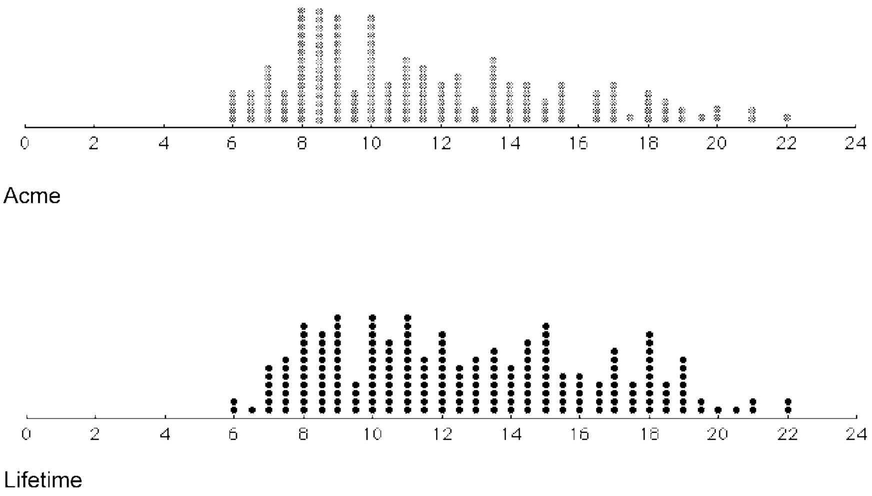


FIGURE 9 Data on ambulance response times.

when the teacher failed to address these issues, and in these instances they refused to conduct an analysis (Tzou, 2000).

As the students worked in pairs at the computers to analyze the data, the teacher and other members of the research team monitored their activity to make decisions about the most effective way to structure the subsequent whole-class discussion. Both these initial observations and a retrospective analysis indicate there was also no regression in the students' ways of organizing data. The majority of the students reasoned multiplicatively about the datasets as they developed their arguments. This observation is consistent with the issues that emerged as topics of conversation in the whole-class discussion. For example, the second pair of students who shared their analysis during this discussion explained that they had partitioned the datasets using the Two Equal Groups option and reasoned with the data points hidden as shown in Figure 10. As they talked, the teacher wrote "½" in each of the intervals to clarify their approach. The students then explained that the lower 50% of the response time data on the upper graph (Acme Ambulance Company) fell in the range of 6 to 10 min, whereas the lower 50% of the response time data on the lower graph (Lifeline Ambulance Company) fell in the range of 6 to 12 min. This, in their judgment, justified recommending that the school district choose Acme Ambulance Company. Crucially, this argument appeared to be constituted by the classroom community as legitimate and

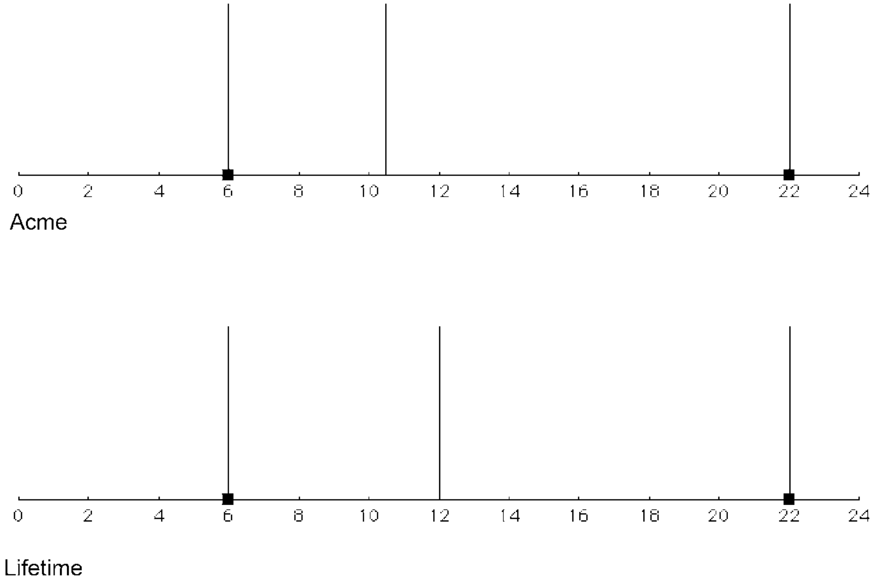


FIGURE 10 Ambulance response time data partitioned into Two Equal Groups with data hidden.

beyond question. Our conjecture that reasoning of this type had been established as normative by the classroom community proved viable as we analyzed the remainder of the data corpus in that we did not find any incidents in which either such reasoning was constituted as illegitimate or reasoning that violated this proposed norm was constituted as legitimate.

The next pair of students who presented their argument built on this analysis by explaining that they had reasoned about the data in a similar manner, but had partitioned the datasets using the Four Equal Groups option (see Figure 11). We note in passing that their reference to the similarity between their own and the previous analysis itself contributed to the constitution of the prior analysis as legitimate. They then went on to explain that they had noticed that if the datasets are partitioned at 13 min, only 50% of the Lifeline ambulances are below this value compared to almost 75% of the Acme ambulances. This explanation was again constituted as legitimate and our examination of subsequent episodes indicated that reasoning in this way about data partitioned into Four Equal Groups had been established as normative.

Taken together, these claims about normative ways of reasoning about univariate data indicate that the conjectures we made about the starting points for the eighth-grade experiment were viable. As a point of clarification, it is worth recalling that these claims have as their focus the classroom microculture that constituted the social context of the students' learning rather than individual students'

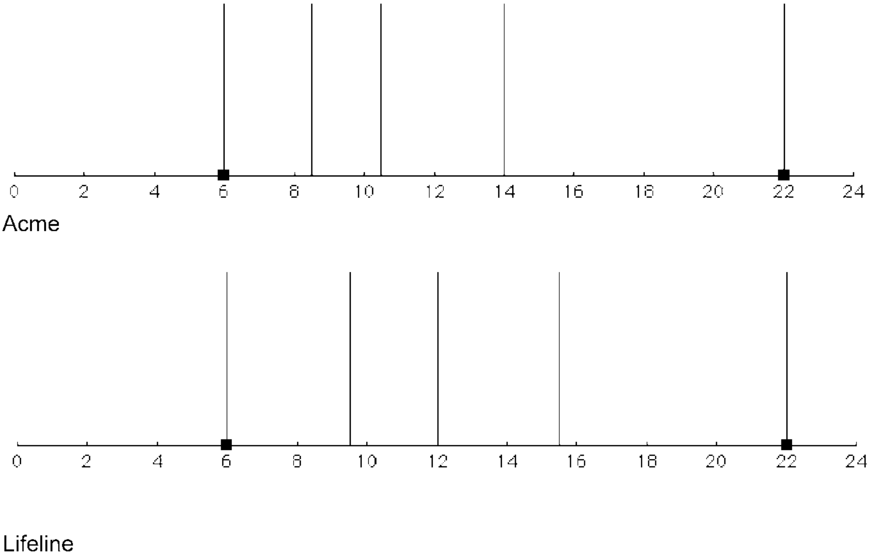


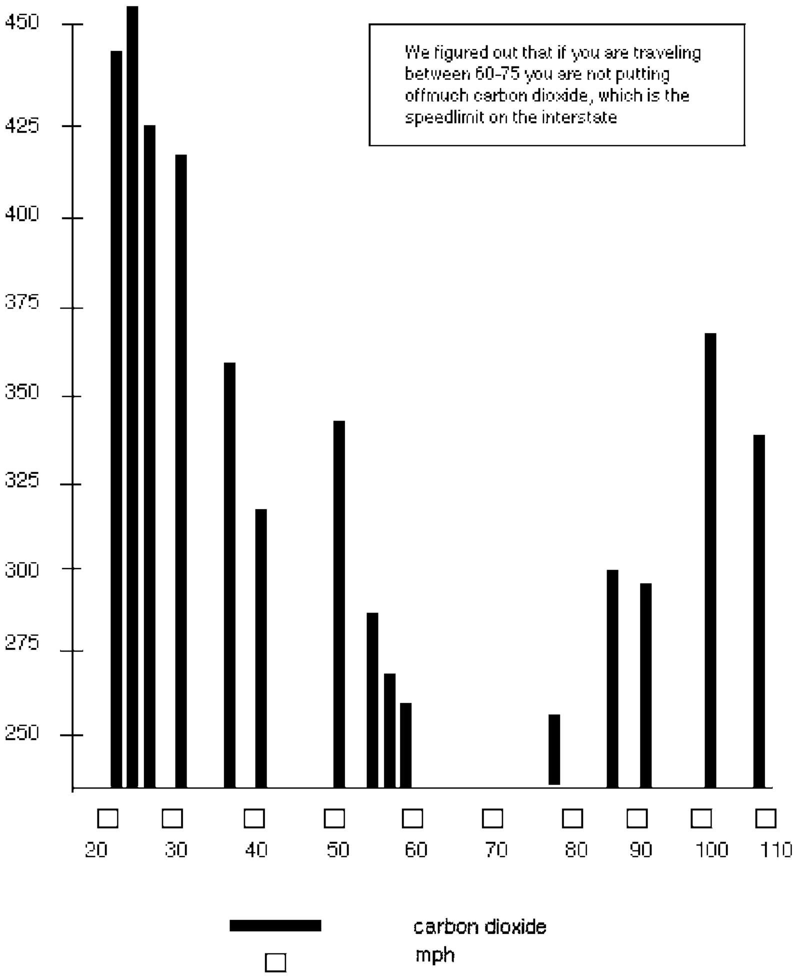
FIGURE 11 Ambulance response time data partitioned into Four Equal Groups with data hidden.

reasoning. As a consequence, we are not claiming that all the students could compare univariate datasets by creating graphs of data partitioned into two or four equal groups or even that they could all interpret such graphs as texts that showed how datasets were distributed without assistance. However, we did observe indications of progress when we focused on individual student's reasoning. At the end of the seventh-grade experiment, almost all the students kept the individual data points visible when they used the minitool to conduct analyses. In contrast, early in the eighth-grade experiment all but a few students began to hide individual data points when conducting their analyses. This was the case when they used the Equal Interval Widths option as well as for the Two Equal Groups and the Four Equal Groups options.

As a final observation, we also note that the normative ways of describing graphs both when the students worked at the computers and during whole-class discussions involved talking about the proportion of data in various intervals without reference to the shape of the data. In this respect, the classroom discourse was consistent with the normative ways of reasoning about Equal Interval Width and Four Equal Groups displays that had been established in the latter part of the seventh-grade experiment. Although we did not realize it at the time, this exclusion of a concern for the underlying shape of data distributions from classroom discourse was to prove significant later in the design experiment.

Inscribing Bivariate Data

The focus of the next phase of the design experiment, which spanned six classroom sessions, was on developing ways of inscribing bivariate data. The first instructional activity dealt with the relation between the speed a car was driven and the amount of carbon dioxide (CO_2) that it emitted. This activity was one of several that were organized around the theme of global warming. During a lengthy discussion of the data creation process, the teacher established with the students that a car was driven for 1 mile at different constant speeds and that the amount of carbon dioxide emitted during each test run was measured by weight in milligrams. The students were then given data that had been generated in this way and were asked to draw a diagram or graph that would enable them to develop a recommendation for the speed limit on interstate highways. Approximately half the students drew double bar graphs as we had anticipated. In addition, a number of these students treated one or both quantities as nominal rather than continuous quantities (i.e., they ordered the measurements of speed and marked them at equal intervals along an axis rather than structuring the axis as a metric space of speed values). In contrast, two pairs of students drew orthogonal axes and plotted the data points as dots. However, one pair put speed on the horizontal axis and carbon dioxide on the vertical axis, whereas the other pair did the reverse. When questioned, the students in these pairs explained that each dot on their graph showed “the speed and [amount of] carbon dioxide together.” This indicates that, for these students, the data consisted of cases for each of which two measurements had been made. The remaining pair of students also constructed orthogonal axes, but then drew bars to indicate the carbon dioxide measurements (see Figure 12). When questioned, these students indicated that they viewed their graph as a bar graph and said that the points at the top of each bar did not show the speed measurements. Thus, although the axes they drew seemed to indicate that they viewed



We figured out that if you are traveling between 60-75 you are not putting offmuch carbon dioxide, which is the speedlimit on the interstate

FIGURE 12 Students’ graph of speed and carbon dioxide emissions.

speed as a continuous quantity, they in fact appeared to treat it nominally. A speed measurement in effect served as the label for a test run during which the amount of carbon dioxide signified by the bar was emitted.

The subsequent whole-class discussion initially focused on the two scatter plots that two pairs of students had produced, but with the axes oriented differently. The teacher first asked about the meaning of the dots on these graphs:

Teacher: So, if I picked any dot what would that tell me? What information would I get from the dot? I'm asking everybody that's in here. What information does that dot give me? What do I know? You got any ideas Mark.

Mark: Do I have any ideas? Yeah. It represents, like lines that go this way, all these numbers and stuff here, the lines that go vertically, like the 53, speed, miles per hour. If it goes this way its', it's …

Teacher: Carbon dioxide.

Mark: Yeah how much carbon dioxide. If you put it together then you get a dot.

Teacher: Brad, did that make sense?

Brad: Yes.

Teacher: So did Mark's explanation make sense? People have questions for him? Kyle?

Kyle: I agree with Mark. You've got to have the miles per hour and the CO_2. And when you connect the two together you get this. It's kind of like showing how the difference varies. I can't find out the information if I only have one, like miles per hour.

Teacher: Okay, so Kyle said I can't find out the information if I only know one of these things. Is that fair? I can't find the information by finding one of those things. If I find one of these dots, what could I use this graph to find out about that dot? Wes?

Wes: You could find out, find out the carbon dioxide and the miles per hour.

In the course of this brief exchange, the interpretation of a dot on a scatter plot as signifying a case with two measures appeared to be established as normative. We make this claim for several reasons. First, we could not identify a single incident in the remainder of the data corpus in which a student violated this interpretation. Second, when the teacher later focused the discussion on the graph in which a pair of students had drawn bars to indicate the carbon dioxide measurements (see Figure 12), one of the students who had produced the graph explained that it showed the "same thing" as the scatter plots. He went on to clarify that the speeds were shown by the distance of the bars from the vertical axis. Crucially, his explanation was constituted as legitimate. Third, in a later session, the teacher introduced the third computer minitool and showed students the Dots option. The students all seemed to indicate that they considered the subsequent discussion of the meaning of individual dots to be pointless because they took it as self-evident that each dot signified a case with two measures.

The final indicators concern the ease with which the teacher was able to initiate shifts in the discourse such that the relationship between sets of measures became the topic of conversation. In the case at hand, for example, the following exchange occurred as the teacher and students discussed the scatter plot in which the amount of carbon dioxide emissions was marked on the horizontal axis and speed on the vertical axis. The students who produced this graph had drawn line segments between adjacent dots:

> Val: If these are the speeds like these are the vertical lines, you can see how fast they go. One of the lowest speeds has the highest emission of carbon dioxide. And the fastest speed has a relatively low emission.
>
> Mark: I understand.
>
> Teacher: You do or don't?
>
> Mark: I do.
>
> Teacher: Something to add to that Ben?
>
> Ben: I was gonna say that what the graph shows is that since the order that the lines are in, the order of lowest to highest speeds, it shows you the rise and fall of how much carbon dioxide is let off as you go from a low speed to a high speed.
>
> Teacher: Who understands what Ben just said? Who has a question for Ben if you didn't understand what he said?
>
> Mark: Can you say it slower?
>
> Teacher: Maybe it would help if you came up here and point to the graph.
>
> Ben: (Comes to the front of the classroom). Okay, because the dots are, the order of how they connected the dots are in order from the lowest to the highest speed like they did and not this dot is over there and they're all over the place. It shows you from lowest to high speed the rise and fall of how much carbon dioxide was emitted. So as the speed goes up, it's mostly at the beginning the carbon dioxide goes down.

Ben's explanation was taken as a basis for the remainder of the discussion, which focused on setting an interstate speed limit that minimized carbon dioxide emissions.

Taken together, these indicators that the interpretation of a dot in a scatter plot as signifying a case with two measures was normative imply that the conjectures we developed when formulating a hypothetical learning trajectory appeared to be viable for this initial phase of the experiment. As will become clear, we had far more opportunities to learn in the remaining phases of the experiment. We can also note in passing that a significant aspect of the teacher's role was to initiate shifts in classroom discourse. In the sample episodes, for example, she first framed the topic of conversation as that of the meaning of individual dots before initiating shifts so that it became the relation between the speed and the carbon dioxide measures and then setting the interstate speed limit. Although the students could also initiate such shifts, the teacher was typically able to sanction them if they would not, in her view, be productive with regard to her evolving instructional agenda. The teacher was therefore constituted as a social authority in the classroom (P. Cobb, 1995) in that she was able to control the nature of the discussions in which the students participated.[9] One of her primary obligations was to guide the emergence of mathematically significant issues as topics of conversation by building on the students' contributions. Beyond this, she also attempted to ensure that these issues emerged in such a manner that they built on each other. For example, it was evident that as the students worked in pairs, the task for most of them became to draw a graph rather than to develop a way of inscribing the data that was relevant to the question of setting the interstate speed limit. However, the teacher did not attempt to focus the discussion on this issue until she judged that both the interpretation of dots as signifying cases with two measures and the interpretation of a scatter plot in terms of a relationship between the two sets of measures had become normative. The individual interpretations that the students developed as they participated in the discussion of these latter two issues then served as intellectual resources when the teacher finally initiated a shift to the issue of the interstate speed limit. As a consequence, the discussion could focus on interpolating between data values to identify the speed at which the minimum amount of carbon dioxide would be produced. In the process, the students were afforded the opportunity to reconceptualize the purpose for which they had produced the graphs.

Reducing Scatter Plots to Lines

The third phase of the design experiment, which involved 12 lessons, began with the introduction of the third computer minitool. The teacher first showed the students the four options for structuring data in the minitool (i.e., the Cross, the Grids, Two Equal Groups, and Four Equal Groups). The data used in this introduction were those of carbon dioxide level (particles per million or ppm) and time (years)

[9] In speaking of classroom discussions, we use the term *participation* broadly to include active listening characterized by attempts to understand others' contributions.

for a 22-year period as shown in Figures 5 through 8. Our intent in allowing the students to select from the four options when they conducted analyses was to enable them to organize data in ways that they viewed as reasonable. The initial discussions of the various options progressed smoothly. For example, the teacher established with the students that, in the Four Equal Groups display, 25% of the data for the indicated years were in each of the four regions within a slice. In addition, most of the students used the Grids, the Two Equal Groups, or the Four Equal Groups options, which partition data into vertical slices, when they analyzed datasets at the computers. In addition, almost all the students indicated that they found the task of describing a global relationship between the two quantities relatively routine. For example, the students quickly agreed that the data in Figures 5 through 8 showed that the CO_2 level had increased continuously between 1957 and 1979. Most also indicated that the data display showed "the same thing" when the minimum value of the vertical axis was changed from 300 ppm to 0 ppm despite the dramatic transformation in the visual appearance of the data.

Although these observations were encouraging, it also appeared that the dots within a slice signified nothing more than a configuration of data points. In the case of the data in Figures 5 through 8, for example, a vertical slice did not appear to show the distribution of carbon dioxide measures for a particular time period for any of the students when they used either the Grids or the Four Equal Groups option. Instead, most of the students appeared to read the scatter plot by identifying a global relationship among the measures by "eyeballing" the data and then using either the Grids, the Two Equal Groups, or the Four Equal Groups option to trace a line that fixed this relationship precisely. For example, in a subsequent instructional activity, the students again analyzed data on CO_2 (particles per million) and time (years), but this time for a 40-year period (see Figure 13). In the whole-class discussion of the students' analyses, the students all seemed to assume without question that the purpose was to discern the overall trend in the data. The issue at hand was that of determining the relationship between CO_2 level and time. The students made a range of proposals that all involved using either Grids, Two Equal Groups, or Four Equal Groups displays to trace lines through the configuration of dots on the scatter plot. The discussion continued for some time because the different methods the students described gave rise to several points of controversy (e.g., whether the carbon dioxide level increased more quickly after 1979 and whether it was relatively constant from 1990 to 1999).

The first student to explain his reasoning demonstrated how he had focused on the lower cells of the 10 × 10 Grids display (see Figure 14):

> Brad: OK, on the 3-by-3 [Grid] we couldn't really tell anything so we put it on the 10-by-10 [Grid] and you could tell about right here (points to the graph at about 1970) the data took a very steep turn and started going up.

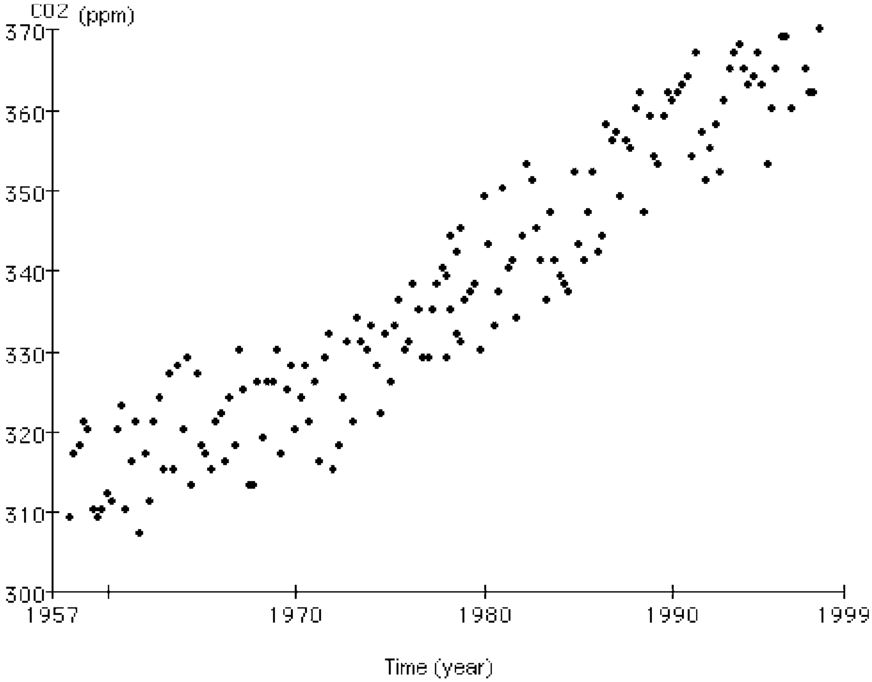


FIGURE 13 Carbon dioxide and years for a 40-year period.

Teacher: So, based on your observations, what's your prediction?

Brad: That it'll keep going up.

Teacher: So what's the "it" in that case, Brad?

Brad: The data.

Teacher: The data? But I thought this was about global warming. You're telling me about…suppose I am a delegate to that conference and you're one of the experts they have brought in to help us decide what we should do about global warming.

Brad: The CO_2 level.

Teacher: OK, the CO_2 level.

This exchange was representative of the entire discussion in that the task for most of the students had become to discern a pattern in a configuration of dots rather than to understand the situation from which the data were generated. This was in contrast to the discussions conducted in both prior and subsequent phases of the design experiment.

As the exchange continued, Brad clarified that he was focusing on the lowest cells in each slice of the Grids display (see Figure 14):

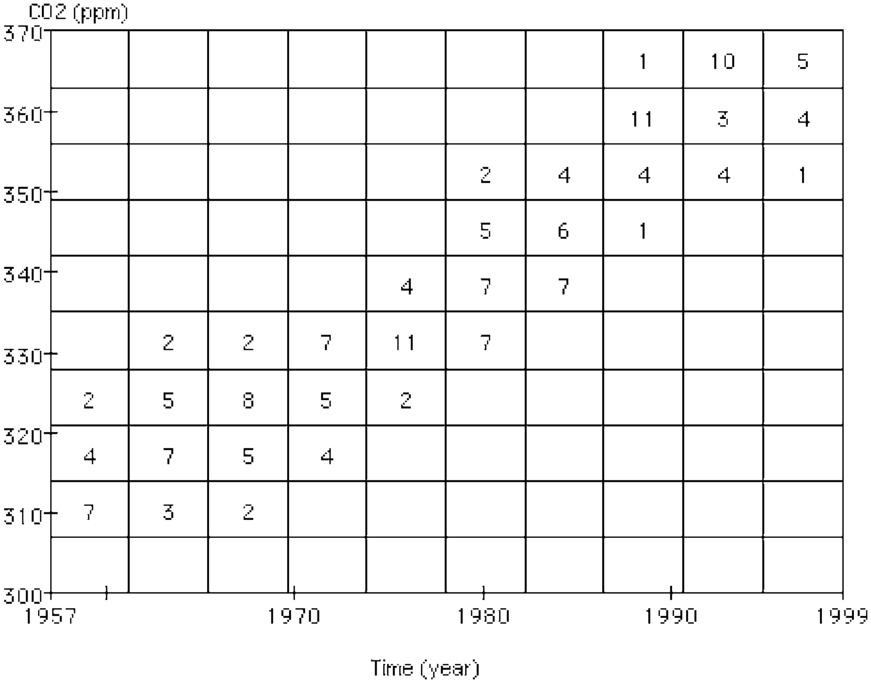


FIGURE 14 Carbon dioxide and years for a 40-year period partitioned using the Grids option.

Janice: I don't really see that the, steep, whatever it is he's talking about.

Brad: OK, right here (runs a finger along the lowest cells of the first three slices), you are going kind of straight looking. And then right here (points to data around 1970 mark), it goes up.

In addition, he explained that he discerned this pattern in the entire configuration of dots:

Teacher: Brad, what would help me: are you talking about in the bottom part of the graph, you talked of the graph, which part of the data are you talking about? Are you talking about all of it or are you talking about…

Brad: This right here (circles the dots in the lower left corner of the graph with both hands).

Several students responded by arguing that several dots did not fit the pattern that Brad claimed to have identified. Mike, Brad's partner, replied by first drawing a line that connected the highest values in each slice of a Two Equal Groups display and a second line that connected the lowest values of each slice. He then sketched a third line approximately midway between the first two, arguing that it slanted up more steeply at about 1979. Mike's attempt to support Brad's argument is of interest because it indicates what he and most of the other students meant by a trend or pattern. He used the Two Equal Groups option to determine the upper and lower boundaries of the configuration of dots and then drew a third line to specify the trend in the entire configuration. It is in this sense that we speak of the students reducing a scatter plot to a line. This is presumably what Brad meant when he gestured with both hands while explaining that he had identified a pattern in the entire configuration of dots. It is important to note that the students did not challenge this general method of attempting to determine the boundaries of the configuration. Instead, they challenged the specific proposals for specifying the boundary. However, because the focus of the discussion was empirical and on whether some of the dots violated a claimed pattern, the students were unable to resolve the differences in their viewpoints. This was in contrast to other phases of the design experiment in which the students were obliged to explain why the way in which they had structured the data was relevant to the question or issue being addressed.

We stress that the conclusions we drew from the exchanges involving Brad and Mike apply both to the remainder of this discussion and to other discussions in this phase of the design experiment. It is also worth noting that the students could, without exception, describe the trends they identified in terms of a relationship among quantities (e.g., CO_2 level and time) when they were pressed to do so by the teacher. When we discussed our observations during research team meetings, we distinguished reasoning that involves reducing a scatter plot to a line from an alternative type of reasoning in which a line is traced to indicate a conjectured relationship of covariation about which the data are distributed. Our inference that scatter plots had not been constituted in the classroom as bivariate distributions implies that most of the students were, in a very real sense, not doing statistics. The generally accepted ways of dealing with bivariate data that seemed to be emerging involved substituting certainty for variation rather than developing ways to manage uncertainty. Clearly, these developments were highly problematic given the overall goals that we had established for the design experiment at the outset.

On the basis of this analysis, we decided to modify the planned instructional sequence in two ways. First, we decided to develop instructional activities in which the students would analyze what we referred to as *stacked* data. These were datasets in which the measures of one quantity were ordered but discontinuous so that the data appeared on a scatter plot as a series of vertical stacks. The example shown in Figure 15 reports the results of an experiment that investigated the effect of alcohol consumption on reaction time. We conjectured that, with the teacher's guidance, datasets of this type might come to be viewed as a series of univariate

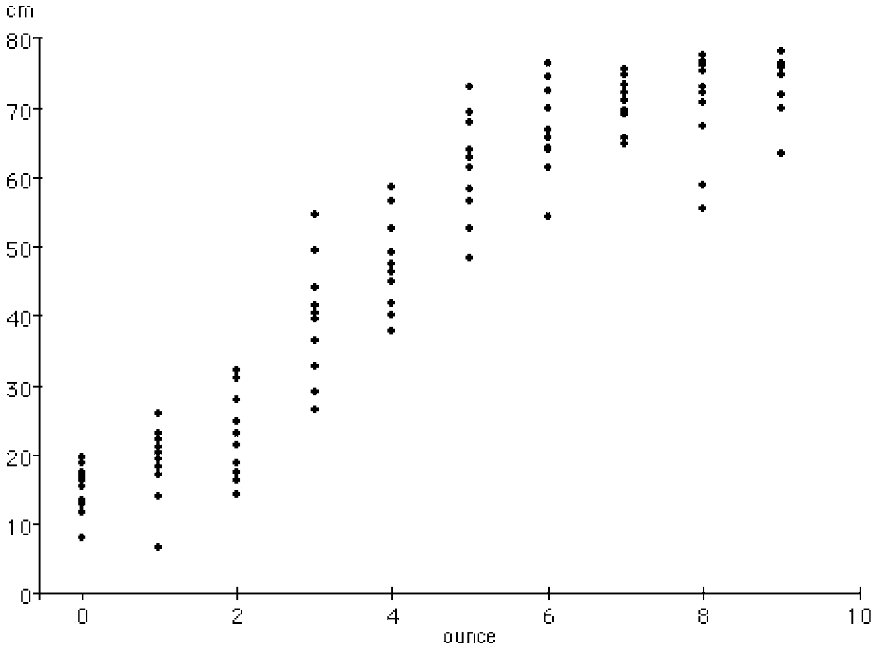


FIGURE 15 Alcohol and reaction time data.

distributions. As some of the dots overlapped, students would have to read how the data were distributed by using Grids or Four Equal Groups displays. However, this appeared to be feasible given the advances made during the first phase of the design experiment when the students used the Equal Interval Widths and the Four Equal Groups options of the second minitool. If the interpretation of data stacks as univariate distributions became normative, we planned to build on it during subsequent discussions that would focus on the distribution of data within slices of scattered data (i.e., a scatter plot).

The second design decision we made was precipitated by the apparent arbitrariness of the lines the students traced through datasets. This indicated that it might be important to engage the students in discussions in which characteristics of univariate data that are relatively stable across samples became an explicit topic of conversation. We reasoned that if the view that the median of a dataset is relatively stable when compared with the extreme values became normative, the students might consider it natural to trace a line roughly through the medians of univariate stacks (or slices) to indicate a relationship of covariation about which the entire bivariate dataset is distributed. To this end, we planned to use stacked data as the basis for discussions in which the teacher would ask the students about the stability of the median and extreme values if an experiment were repeated and another sample were generated. We chose to focus on the median rather than the mean because the students were relatively familiar with it as a consequence of using the Two Equal Groups and the Four Equal Groups options of the second minitool during the seventh-grade experiment.

Negotiating the Median

As will become apparent, we were repeatedly surprised by the students' reasoning during the nine classroom sessions in which we investigated the two revisions we made to the initial learning trajectory. The teacher typically responded to students' unanticipated contributions by asking follow-up questions in an attempt to understand their thinking. As a consequence, the sessions often had the feel of exploratory interviews that were being conducted with an entire class rather than with a single student. There were frequent indications that the students responded on the basis of their mathematical interpretations of the teacher's probes even when sequences of questions could clearly have cued them to an alternative response. It appeared that most of the students had come to view themselves as active collaborators in the design experiment and saw it as their role to help us learn about their thinking and thus improve the instructional activities and the third computer minitool. This impression was confirmed by a number of the students during conversations that a member of the research team conducted with them outside the classroom sessions as part of a separate study. Although we did not consciously attempt to guide the development of a classroom participation structure of this type, we realize with hindsight that the teacher's obvious interest in and responsiveness to the students' contributions contributed to its emergence. The sessions conducted during this phase of the experiment are therefore a particularly rich source of insights into the students' statistical reasoning.

In the first of these sessions, each student generated 10 measures of his or her reaction time using the procedure of grasping a meter ruler that was dropped through an open hand.[10] The teacher began the subsequent whole-class discussion by showing plots of 10 of the students' reaction time measured in centimeters with the medians marked (see Figure 16). She then questioned them about their expectations if another eighth grader's reaction time was measured using the same method. Although some students specified ranges of varying magnitudes, others argued that it depended on the student and that there was not enough information:

[10] This method of measuring reaction time by asking people to grasp a meter stick that was dropped through an open hand was used in several studies that we found in the literature.

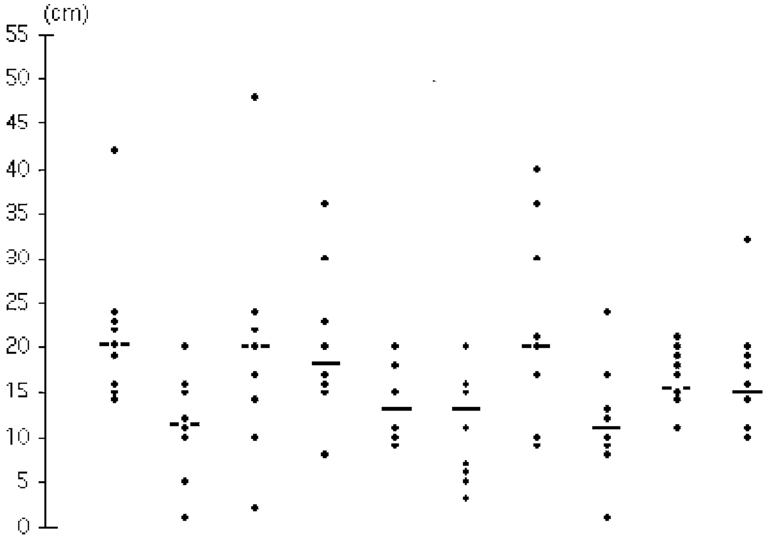


FIGURE 16 Plots of 10 students' reaction time measures with the medians marked.

Teacher: If these [the reaction times of ten students in the class] are typical of eighth graders, what do you think the median of any other eighth grader that did this might be? What might you say about the median?

Brad: Twenty.

Teacher: The median would be about twenty. Suzanne?

Suzanne: I'll say about fifteen and twenty.

Teacher: So Suzanne is saying that in a range of between fifteen and twenty the median would lie. Val?

Val: I don't think you can really tell.

Teacher: You don't think, why not?

Val: Because it's like you don't really have like enough information, I mean, you have information about us but you don't have information about the rest of the students. I mean you just say any eighth grader it could be anything, it could be just up at twenty, it could be at thirty, it could be at ten, it could be anywhere…

In the course of a long exchange with the teacher, the next student to explain her reasoning supported Val's argument:

Sinae: OK, so, OK, if you look at every one of the people, if you look at all of us individually we all vary in so many places, it's like some of us can't even catch the stick (inaudible) fifty, you know stuff like that, while some of us you know can catch it within one centimeter, stuff like that, so I mean it is impossible to say how that person is going to do because they can be like the first person or they could be like second person, or they could be like the third person, fourth, fifth, sixth…

Teacher: Yeah, but if I look at your medians, Sinae, to me they don't seem to be all over the place.

Sinae: I know but…

Teacher: (Interrupts) So I am asking you to talk about the next person's median, not all of their scores.

Sinae: OK, but that person is going to vary, that person, I mean, we are not alike.

The arguments that Val and Sinae proposed appeared to involve what Konold (1989) called an outcome approach in that the task for them was to predict the actual reaction time measure for an unknown eighth grader rather than to make a probabilistic inference.

In contrast to these responses, questions about the reaction times of another group of 10 eighth graders led to little discussion:

Teacher: If ten more eighth graders did this, ten more eighth graders, what would you think about the range of their medians, if ten more did it. Janice?

Janice: I think that basically they will be the same as that [the reaction time data for the students in the class] because like we are all different, like those students up there are all different and they all had different results but I think it'll be the same cause, if you randomly pick ten, some will be slow, some will be extremely fast like, you know, like some of us.

Teacher: Right.

Kim: I would like to say.

Teacher: All right, Kim.

Kim: Eleven to 21 will be somewhere in there or maybe a little higher than that or a little lower than that, but that is where most of us are, so…

Teacher: So you think that the next ten might be in that same, in that same range.

The important aspect of Janice's argument was her assumption that some of the new group of students would be fast and some would be slow, as was the case with the students in the class. We viewed this argument, which was treated as legitimate, as encouraging in that it indicated a possible concern for the way in which the data were distributed.

The datasets the students analyzed in subsequent sessions were organized around the theme of driving safety and accident rates. The first stacked dataset that the teacher introduced involved reaction time data stacked at 10-year age intervals (see Figure 17). During the discussion of the data creation process, the teacher clarified that the 10 data points for each age level had been generated by following the same procedure that the students had used to measure their reaction times (i.e., each data point was the median of 10 measurements for one individual). As was the case during the discussion of previous instructional activities, a general description of an overall trend in the data quickly became accepted. Further, the observation that reaction time becomes more consistent with age was legitimized relatively quickly (i.e., if the outlier in the 50-year-old stack is ignored, the spread of the stacks tends to decrease with age):

> Suzanne: I just wanted to add on to Shane's (comment) that from 20 to whatever that, 90, as they get older, not only does the range ascend, but the range gets smaller.

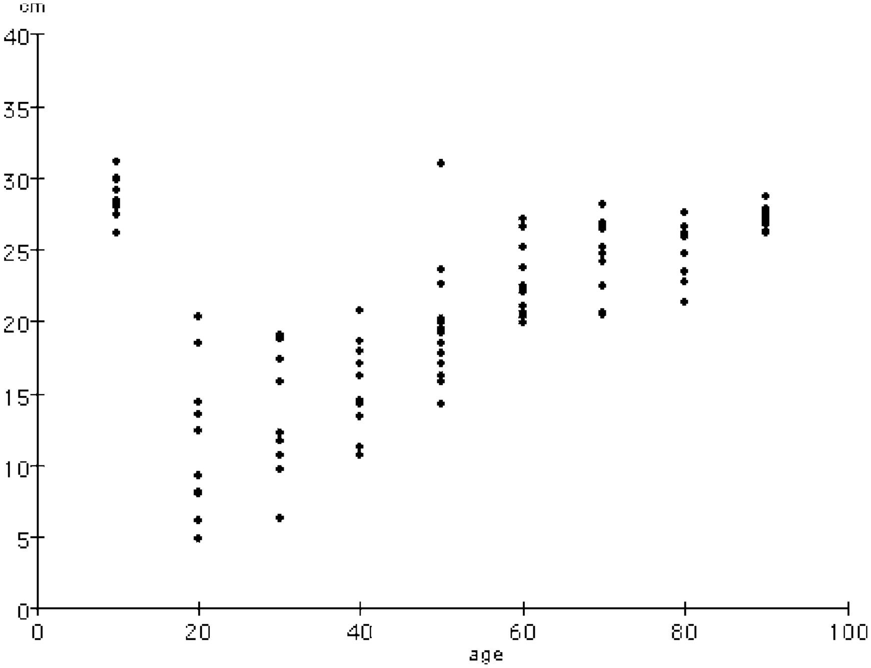


FIGURE 17 Stacked age and reaction time data.

Teacher: So there are two things that I keep hearing you refer to. You're saying that as we look at older groups of people, that the range got higher. (inaudible) And also the range got tighter. Is that correct? Is that what I'm hearing you say? Val.

Val: Didn't you just like cancel yourself out by saying that the range got higher and the range got smaller? I don't understand.

Teacher: Brad help me out.

Brad: Okay. What they mean is the range of the amount of centimeters that it took to catch it went higher, but the range of the dots from the lowest amount to the highest amount. OK, look.

Val: Isn't that the same?

Brad: No. From here to here it went up (points across ages), but from this dot and this dot of the ages it got smaller (points to extremes within the age).

Val: The first range got higher than what? I don't... You just said from here on the 20-year-olds and here on the 90-year-olds it got higher. Higher than what?

Brad: Work with me, okay? The amount of centimeters that it took for them to catch it went higher as the age went up, basically.

Teacher: OK. As people got older, it took more centimeters before they could catch it.

Val: OK, I understand it, never mind.

One student did raise the issue of whether the outlier should be excluded. However, her concern appeared to be that a line traced through the highest values of the stacks would be inconsistent with the overall trend that had been described if it were included. In response, another student recommended focusing on the medians of the stacks because this made it easier to see the "steps":

Kim: If you use the median it's much easier to see if they're going up, down, up, instead of going uh, uh, uh,

Teacher: If we use the median?

Kim: Instead of using the top parts. And you just skip that top part in that 50 one or whatever.

Teacher: So Kim thinks we should use the median.

In making this proposal, Kim suggested a method for reducing the scatter plot to a line that ignored the variability in the data. Contrary to our rationale for introducing stacked data, there was no indication in the course of this discussion that any of the students interpreted the stacks as univariate distributions. Instead, the stacks appeared to be constituted in public discourse as collections of data points that occupied intervals bounded by their highest and lowest values.

We were surprised by these observations, given our assumption that the constitution of the stacks as univariate distributions would involve little if any modification of the normative ways of reasoning that had been established when the students used the second minitool. In an attempt to better understand their reasoning, we decided to probe their views about the behavior of the median if new data were generated. To this end, the teacher began a discussion of the same age–reaction time datasets in the next classroom session by first reminding the students that 10 individuals' reaction times had been measured at each age level. She then focused on the data stack for 70-year-olds and asked the students how the median might change if they were to measure the reaction times of 10 additional 70-year-olds. During the exchange, most of the students indicated that they expected the medians of the two samples to be close, but not exactly the same. However, some of these same students argued that the median would be either lower or higher for a new sample of fifty 70-year-olds. Further, all but one of the students who thought the medians would be about the same gave empirical arguments (e.g., because the data for the 70-year-olds are "close together," the median is "predictable but not exactly"). The student who gave a nonempirical explanation argued as follows:

> Ben: I think it be about the same.
>
> Teacher: Why?
>
> Ben: Because if it's the same for the next 10 people why shouldn't it be the same for the next 10 and the next 10 and the next 10 and the next 10?

The teacher subsequently revoiced Ben's argument in some detail, but the other students were not convinced. It appeared that many of them did not view his argument as being relevant to the question of predicting the median of fifty 70-year-olds. Instead, it seemed that these students viewed a data stack as an amorphous collection of data points located within a particular interval rather than as a distribution.

The teacher continued to explore the behavior of the median with the students in the next classroom session when she introduced data generated during an experiment that investigated the effects of alcohol consumption on reaction time with 20 data values at each alcohol level (see Figure 15). During the initial discussion of these data, it became apparent that almost all the students focused on the highest values of each stack when they described a trend. Once again, there was no indication that any of the students interpreted the stacks as univariate distributions. The teacher next clarified that the legal driving limit in their state was approximately 2 oz of alcohol, which corresponded to two alcoholic drinks of either beer or wine. She then showed the students the reaction time data for only the 0-oz and 2-oz levels and asked them whether the two-drink limit was justified (see Figure 18). Our rationale for posing this task was that this comparison of two data stacks corresponded reasonably closely to the types of instructional activities the students had completed when they used the second minitool. We therefore conjectured that some of the students might view the two stacks as data distributions. However, it became clear almost immediately that most of the students viewed the stacks as collections of data points that occupied an interval rather than as distributions.

In response, the teacher initiated a shift in the discussion by marking the medians of the two stacks. She then asked the students how many reaction time measures they would expect to be above and below the marked medians if the experiment were repeated with 20 people at each alcohol level. The teacher and students quickly established in the ensuing exchange that about half the data points would be above and half below the medians, and the new medians would be close to, but not exactly the same as, the marked medians. We viewed this as an advance when compared to a previous exchange, in which a number of the students indicated that they expected the median to change significantly if the size of the sample were increased. The teacher's questions appeared to have oriented the students to consider where the data

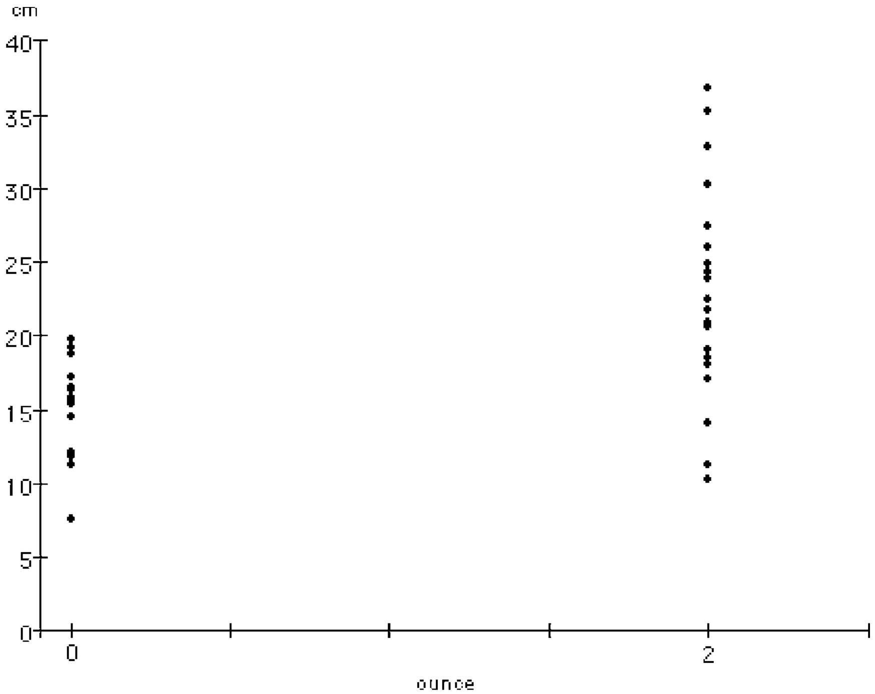


FIGURE 18 Reaction time Data for 0-oz and 2-oz levels of alcohol consumption.

points might be located in relation to the median. This contrasted with the previous exchange, in which most of the students seemed to view the median merely as a point somewhere in the interval occupied by a data stack.

In the next classroom session, the teacher continued to focus the discussion on the relation between data and the median by asking the students to imagine that 20 people's reaction times had been measured after they had consumed 1 oz of alcohol. She then asked the students to suggest reasonable data values if the median was 25 cm. Several students commented that they found this task challenging because they did not know the range. These responses seemed to indicate that the median was for them simply a value somewhere in the interval occupied by a dataset. A second group of students responded by specifying ranges, presumably by drawing on their familiarity with reaction time data (i.e., they expected the median to be approximately in the middle of the range). In contrast, a third group of students appeared to reason in terms of the relation between data and the median when they argued that half the data would be greater than 25 cm and half would be less.

The teacher next plotted a stack of data values with a median of 25 cm and the students indicated that they viewed this dataset as plausible. The teacher then posed a series of questions that explored how the data points would have to be changed to change either the range or the median by specified amounts. In the course of this discussion, the teacher and the students established that the range was sensitive to changes in a small number of data points and that plausible datasets with the same median could have substantially different ranges. They also established that, in contrast, a relatively large number of data points had to be changed to change the median significantly.

Against this background, the teacher then asked the students whether it was more informative to know the range of a dataset or its median. All but one student indicated that knowing the range was more informative because the median was "just one number." We were initially surprised by these responses because we had assumed that the students had reflected on the relation between data and the median as they had participated in the immediately prior exchange. Our intent was that the view of the median as characteristic of a dataset that depended on how the data were distributed might become normative. However, it appeared that, for the majority of the students, the discussion had involved an empirical demonstration of the process of manipulating data points to change the range or median.

The lone dissenter who said that the median was more informative argued that "the median tells you where the majority [of the data] is." However, the other students rejected her argument even though the teacher repeated it by clarifying that the reaction time data were tightly clustered in a portion of the range that centered on the median. This argument was nonetheless significant from our point of view in that it suggested how we might be able to further revise the conjectured learning trajectory. As we noted when we discussed the starting points for the design experiment, the students had used the term majority frequently when they had discussed analyses that they had conducted using the second minitool. This term was typically used to indicate a relatively large proportion of a dataset that was located in a particular interval. We also noted that the origin of this term could be traced to the notion of a hill, which referred to the shape of a univariate distribution and indicated an interval where the majority of the data were clustered. The contrast between the dissenting student's reference to the majority and the explanations of the other students alerted us to the possibility that the discussions in recent classroom sessions might have been completely disconnected from the normative meanings established when the students used the second minitool to analyze univariate data. This possibility would, of course, account for the observations that we had found surprising in these sessions.

In hindsight, it is apparent that we had assumed that the median is a relatively unproblematic notion and had failed to realize that there could be two distinct meanings for the term that are grounded in differing types of activity. The first of these meanings derives from the activity of "finding the median" by manipulating individual data points (e.g., dividing a dataset in half, finding the "middle number," etc.). On reflection, it is apparent that most of the students had reasoned about medians from this perspective in recent classroom sessions. The second type of meaning derives from the activity of viewing univariate datasets as having shape according to how the data are distributed. In this case, the median can indicate a feature of the shape of a data distribution (e.g., a hill where the majority of the data are clustered in the case of reaction time data) as well as a partition of the dataset in half or into two equal groups.

We drew on this distinction to revise the conjectured learning trajectory for the remainder of the design experiment. The first step involved investigating whether the interpretation of data stacks as having shape might become normative. We conjectured that data stacks would then be constituted as distributions rather than collections of data points. This might then make it possible for a display of data stacks to be constituted as a bivariate distribution (i.e., a distribution of univariate distributions). If the revised learning trajectory proved viable to this point, we next planned to investigate the transition from stacked data to scatter plots. We anticipated that a key step in this transition would involve interpreting the slices of Grids and Four Equal Groups displays as univariate distributions.

Reading Stacks and Slices as Distributions

Our investigation of the revised learning trajectory spanned the final six sessions of the design experiment. As an initial assessment of the feasibility of the trajectory, a member of the research team questioned several of the students as they worked at computers individually or in pairs to analyze stacks of 50 reaction time measures for the 0-oz, 1-oz, and 2-oz alcohol levels. It quickly became apparent that these students could readily interpret these data stacks in terms of shape when either the Grids option or the Four Equal Groups option was used (see Figure 19). This was indicated by the ease with which they traced a hill-like shape with a finger on the computer screen. In the case of Four Equal Groups, for example, the students said that the relatively small range of the middle 50% of the data (i.e., the second and third quartiles) indicated a hill. They also said that they expected the shapes of the data stacks to be similar if the data creation process were repeated.

The teacher attempted to capitalize on these observations during the subsequent discussion of the students' analyses. The first student to explain his and his partner's reasoning said that they had used the 10 × 10 Grid option (see Figure 19a) and asked the teacher to hide the data (see Figure 20). He then explained that it was not possible to compare the stacks by looking only at the cells with the highest counts because a lot of the data were in other cells as well:

Mike: You see if you look carefully at it (points to the data stack for 0 oz of alcohol in the Grids display) [see Figure 20], um, there's not really much of the consistency. Because there is also a large amount of data points here (points to a cell with 16 data points). So, you can't just decide on the set [of 17 data points] in here because there are 17 data points in this square and there are also 16 in this square.

Teacher: OK, I'm gonna say what I think you said, and then you tell me if I'm right. So you're saying that you would need to look at maybe this interval (points to the cell containing 17 data points) because it has 17 data points in it and this one (points to one of the cells containing 16 data points) because you are looking at where the majority of the data is clustered here and that would be important because you have all the way across (points to the other two data stacks), right?

In restating this explanation, the teacher attributed to Mike and his partner the intention of using the Grids to find where the majority of the data were located in each stack. She then asked the students if they remembered using the second minitool. All indicated that they did and one student sketched a horizontal axis with a curved line above it to signify the shape of a univariate dataset. Pointing to his drawing, the teacher reminded the students that they had previously referred to this shape as a hill and asked if they could tell from the drawing where the data were bunched up. All indicated that they could and most appeared to view the

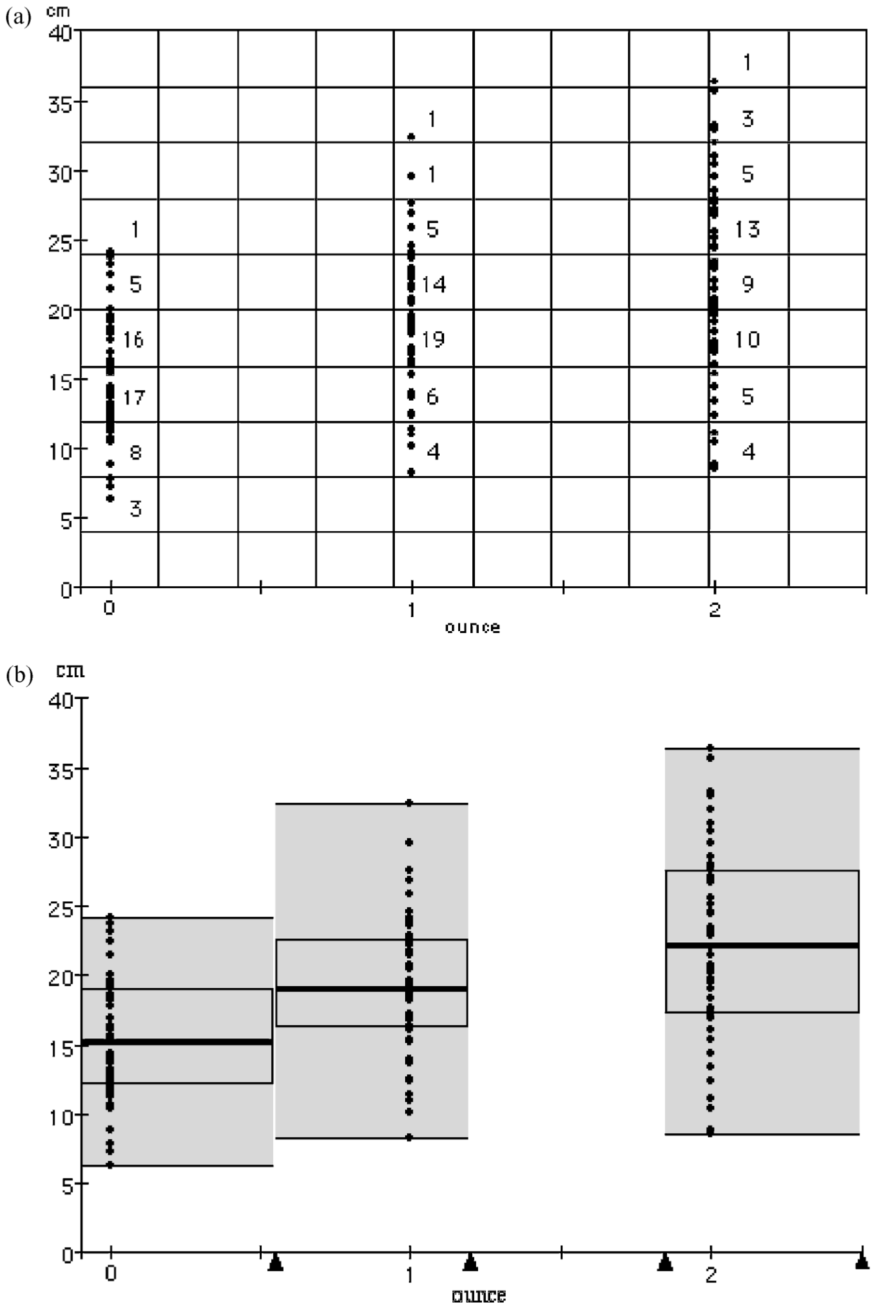


FIGURE 19 (a) Alcohol and reaction time data for 0-oz, 1-oz, and 2-oz levels organized using the Grids option. (b) Alcohol and reaction time data for 0-oz, 1-oz, and 2-oz levels organized using the Four Equal Groups option.

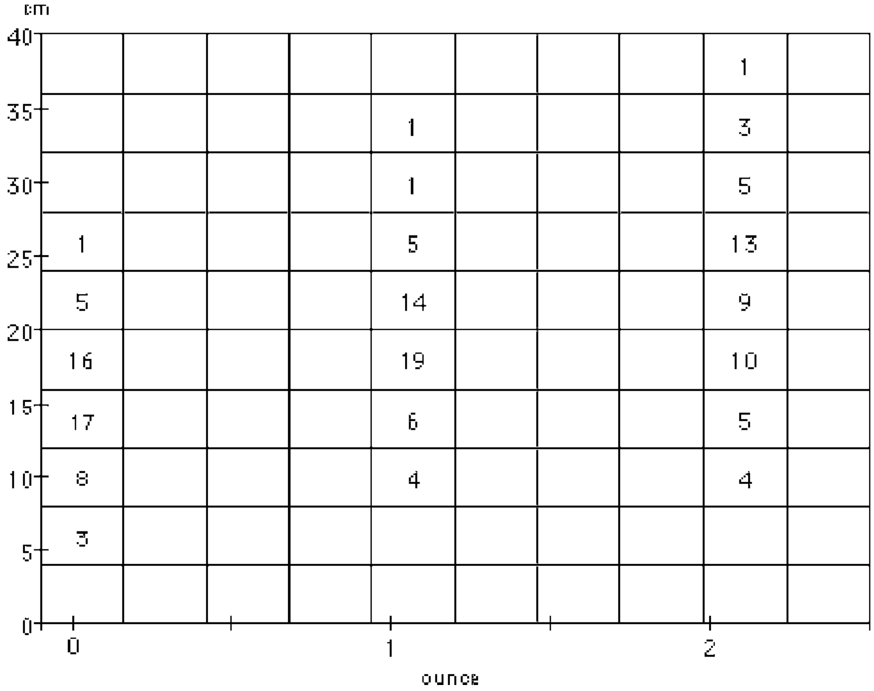


FIGURE 20 Alcohol and reaction time data organized using the Grids option with data hidden.

response to the teacher's questions as self-evident. The teacher then returned to the 10 × 10 Grid display with the data hidden and the discussion proceeded smoothly as she and the students sketched the shape of each data stack. Mark, for example, made the following comments as the teacher sketched the shape of the data stack for 2 oz of alcohol:

Teacher: So if I did the same thing here (points to the data stack for 2 oz of alcohol), I'm just gonna use this for my baseline (runs a finger along the data stack)…that one kind of goes like that…and this one kind of does like that (sketches the shape).

Mark: It's a sideways hill, man.

Teacher: It is a sideways hill. Now, let me ask you, what if we looked at it, anybody have a question about this, does this make sense?

Mark: I can see why you have two little bunches, but you know, it's like, irrelevant.

Teacher: Yeah, because I'm trying to show you where the data's bunched up.

Mark: Trying to show where the like 10 and the 9 and the 13…
Teacher: Right, exactly Mark.

The teacher next showed the students the Four Equal Groups display (see Figure 19b) projected onto the shapes she had sketched and asked the students what this told them. The students' initial comments focused on either the median or the percentage of data in various parts of the display. This type of discourse is reminiscent of that which had occurred when the students used the Four Equal Groups option of the second minitool during the first phase of the design experiment. As we noted, normative ways of talking about Four Equal Groups displays involved describing the proportion of univariate datasets in various intervals rather than their shape. However, a shift occurred in the discourse when the teacher pointed to the 1-oz stack and asked how much of the data was in the middle two sections (i.e., the second and third quartiles):

Teacher: How much of the data is in this one and this one together (points to the second and third quartiles of the 1-oz data stack)?
Students: 50 [percent].
Teacher: Well how come this one (points to the first quartile of the 1-oz data stack) is so much wider than these two put together?
Kim: Because they're spaced out.
Teacher: Brad.
Brad: Because right there in that little middle spot there is a whole lot of data points. That's where the hill is because that's, they're more clumped together in the middle than up top and down bottom there is not as many.
Teacher: Which is what you said, Kim. They are more spread out. So, that's why my hill is happening right there because that's where they're bunched up. Same thing here, right?

The teacher next asked the students if they could predict what the data might look like if the reaction times of 50 people who had consumed 1/2 oz of alcohol were measured. As she and the students interpolated from the 0-oz and 1-oz stacks, it became established almost immediately that the Four Equal Groups display showed the shape of the data stacks. Further, the students indicated without dissention that the shapes of the data stacks would be similar if the experiment were repeated with samples of 50 people at each alcohol level, and with samples of 100 people.

We interpreted these observations as indicating that the revised learning trajectory might be viable. It might also seem reasonable to conclude that the interpretation of the Grids and the Four Equal Groups displays of stacked data in terms of the shape of univariate data distributions had been established as normative. However, an exchange that occurred at the end of this class discussion calls this latter conjecture into question. Immediately after the students had agreed that the shape of a data stack was predictable (i.e., it would be similar if the data creation process were repeated), the teacher asked them whether the median or the high and the low values of a data stack were more predictable. As in previous class sessions, several students argued in favor of the high and the low values. One reasoned, for example, that the median "can be anywhere in there." This suggests that the teacher's question had oriented these students to view a data stack as a collection of data points occupying an interval rather than a data distribution that had shape. Several other students who contributed to this exchange indicated that they thought the median was more predictable, perhaps because they saw it as indicating the location of a relatively stable hill.

In light of this exchange, we modified our conjecture about reading the Grids and the Four Equal Groups displays in terms of shape by adding the qualification that this interpretation appeared to be normative only when reasoning about data stacks, but not when reasoning about individual points within a stack. We stress that it was not the diversity in individual students' views of the predictability of the median and extreme values that led us to make this qualification. Instead, it was that none of the students' interpretations were established as more legitimate than those of other students.

One of the students who argued that the median was more predicable than the extreme values developed a relatively sophisticated explanation. His argument gives insight into the demands of viewing the median as a characteristic of an entire dataset rather than as a point in the range when the interpretation is not grounded in an image of the shape of the data distribution. He argued that if the experiment were repeated and one person who was not alcohol-tolerant took 60 cm to catch the meter stick, the range would move a lot, but the median would move only a little. The teacher asked the other students for their views on this argument, but elicited little reaction. Crucially, in articulating his argument, this student had reasoned about the relation between changes in individual data values and changes in the median. He had proposed several other arguments in prior classroom discussions that reflected this relatively sophisticated view of the relation between data and the median. From conversations conducted outside the classroom sessions as part of a separate study, we found that most of the other students viewed his contributions as irrelevant and as disrupting the flow of discussions. Presumably, in situations where we failed to support a focus on the shape of a distribution, for these students the median was a value within the range that was identified by carrying out a calculational process that they had previously been taught. However, for the student whose contributions were viewed as irrelevant, the median was not a value in the range on a par with individual data values. Instead, it seemed that he viewed this calculational process as not merely one of finding a numerical value, but also as one of structuring a dataset by partitioning it.

As a consequence, he could treat the median as a structural characteristic of a dataset that could be reasoned about in relation to the entire dataset.[11] The repeated miscommunications between this student and the other students highlight the epistemological gulf between their ways of participating in classroom activities when their reasoning was not anchored in images of the shape of data distributions.

The next analysis that the students conducted involved a comparison of two sets of stacked data of salary against years of education, one for men and the other for women (see Figure 21). There were 20 data points in each of the six stacks in each dataset. We developed this comparison activity with the expectation that general descriptions of global trends (e.g., salary increases with years of education) might be seen as inadequate, given the issue of investigating possible inequities. When the students worked at the computers, none interpreted the stacks in terms of shape on their own initiative. However, most could do so readily when a member of the research team asked them to trace the shape of a stack organized into Four Equal Groups. Furthermore, this brief intervention was sufficient to reorient several students' analyses. For example, one student explained that there was a lot to think about because the data involved men versus women and salary versus years. He and his partner then organized both datasets using the Four Equal Groups option (see Figure 22) and began to find the extreme values of data stacks "to get the overall range." At this point, the students appeared to view a stack as a collection of data points that occupied a particular interval. When the researcher asked the student who had spoken previously if he could trace the shape of a stack, he did so immediately and explained that the hill was where the majority was located. He then pointed to the high values in a stack and added that there were only a few "up there," implying that the data were spread out in that part of the range. The crucial feature of this exchange from our point of view was that what had initially appeared to the student as an almost overwhelming mass of data points seemed to have structure once he traced the shape of one stack.

During the whole-class discussion of the students' analyses, reading the shape of data stacks from displays of the Two Equal Groups, the Four Equal Groups, and the Grids again became established as normative. In addition to speaking of hills and the majority, one pair of students referred to the "humps" where most of the data were and another spoke of the "cluster." A third student who had used the Four Equal Groups option referred to the data being "squished up" (see Figure 22):

[11] In Sfard's (1991) terms, this student's view of the median as he participated in these discussions was structural, whereas those of most of the other students were operational. This student's explanation also clarifies why sampling distribution has proved to be such a challenging notion even for college students (G. W. Cobb & Moore, 1997). It is only when students can reason about the median in structural terms that they can begin to envision the potential ways in which the medians of samples might be distributed.

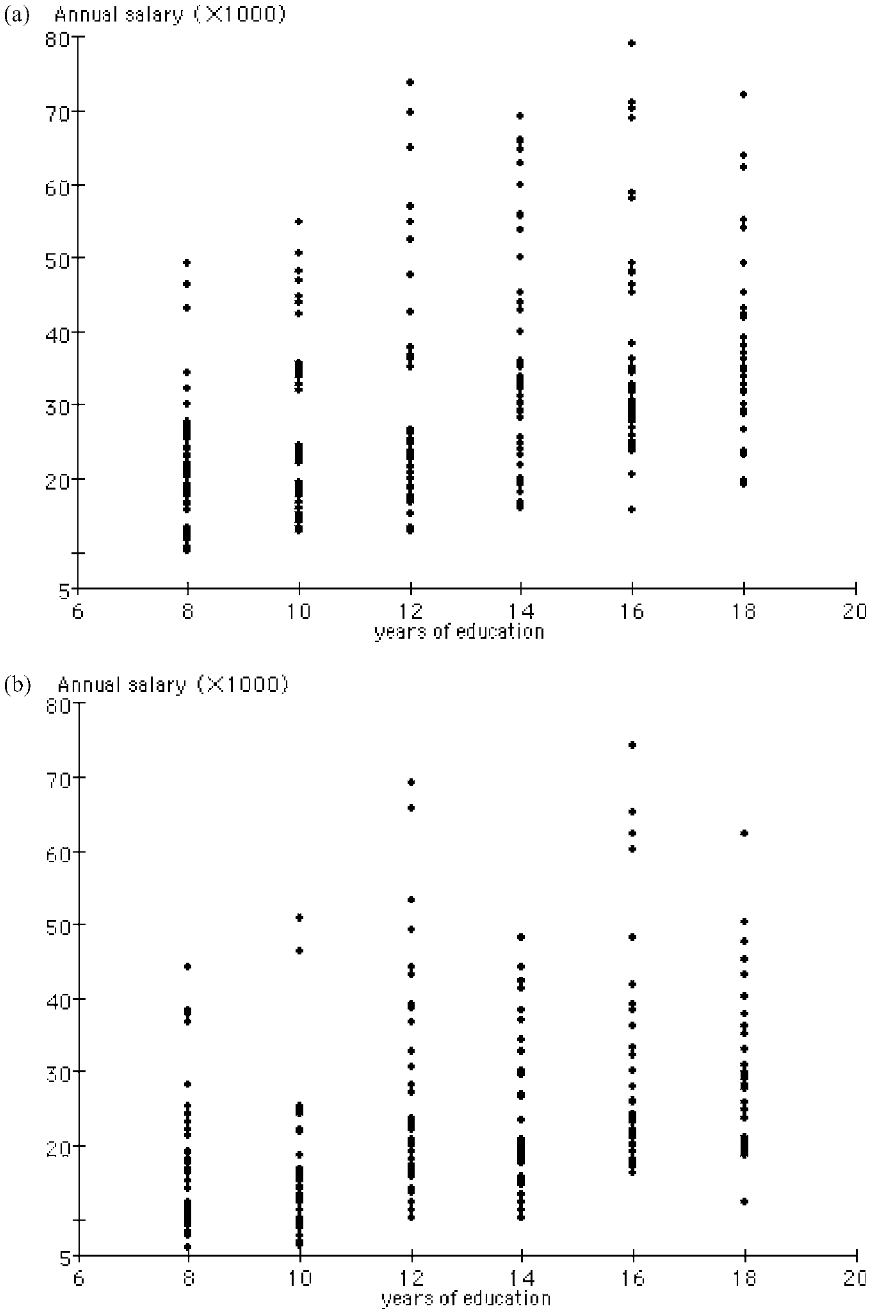


FIGURE 21 Years of education and salary data for (a) men and (b) women.

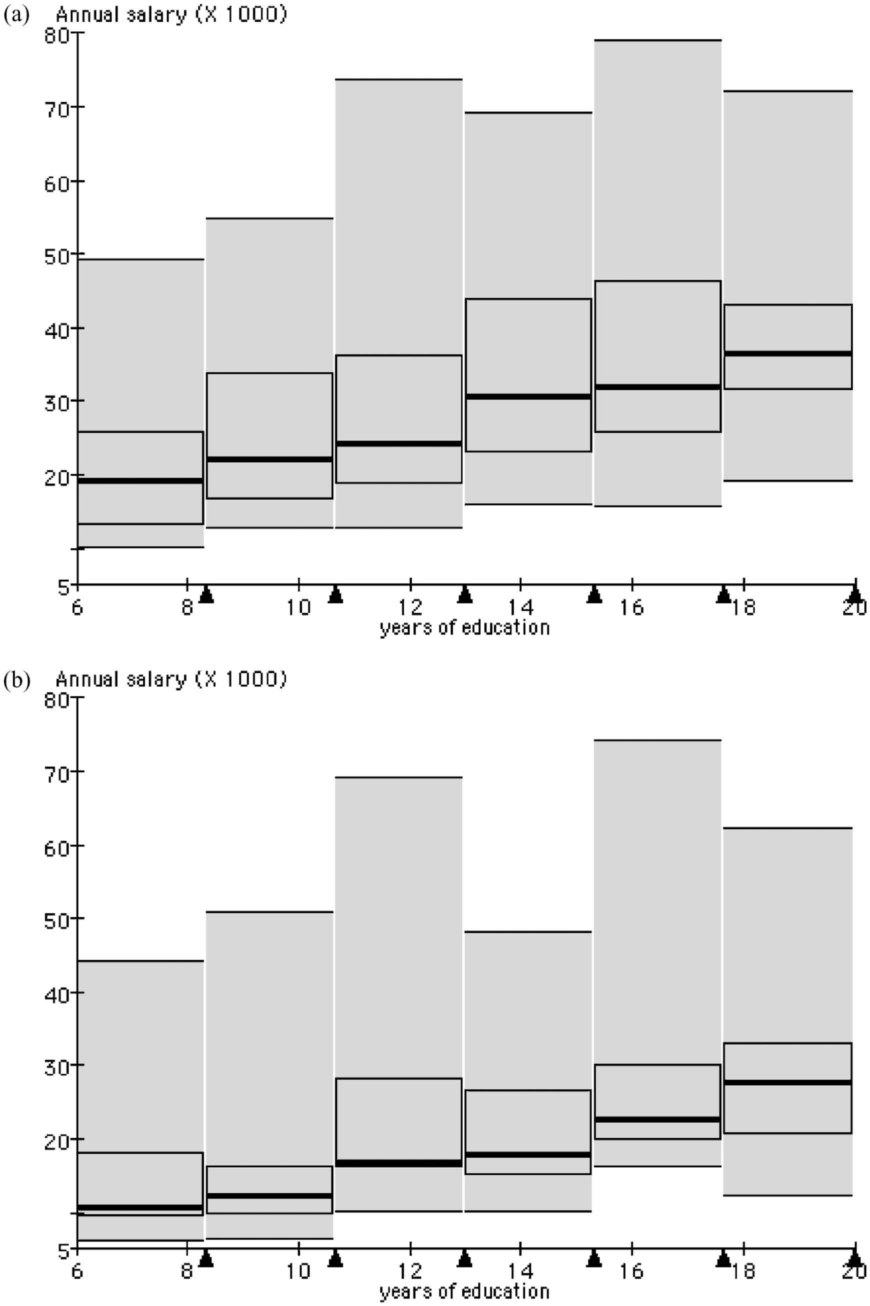


FIGURE 22 Years of education and salary data for (a) men and (b) women organized using the Four Equal Groups option.

Suzanne: (Speaks to the student who is operating the computer projection system) Six-by-four equal groups please. OK, what I did is I looked at the median in each one, and what is this, is this the men's, yes, like the median of this one (points to the data stack for men with 8 years of education) is like $20,000 and the median of the women's is like $13,000 (points to the data stack for women with 8 years of education), so it's about 6 or 7 thousand dollars difference. And like the hump or hill whatever is in like the middle 50 percent. And for the men's the median is 20,000 so the lower 10 data points are like below that and higher 10 data points are above that and you can tell that like the lower 10 are like all squished up together but the higher 10 are more spaced out. And the same thing with the women's only the lower 10 is more squished up than the men's. And like the extremes are even like lower there than it is there (points to the data stacks for 8 years of education on the two graphs) and lower there than it is there (points to the data stacks for 10 years of education on the two graphs). So the men's, the men make more any way.

Teacher: Questions for Suzanne about her way? Val…

Val: So you basically used the median. You basically did the same thing they [a previous group] did by like comparing the medians and the grouping and stuff?

Suzanne: Yeah.

Teacher: Val, do you know why they chose the medians?

Val: 'Cause they wanted to do the humps, I guess. It's just easier to get like one little area…

Brad: That's where most of it is.

Val: It's easier to get an area to compare when you have such, just one set area to look at, I guess.

We take this exchange as relatively strong evidence that the reading of data displays in terms of shape or distribution was normative. Val seemed to view a data stack as a collection of data points that occupied an interval rather than as a distribution. For her, focusing on the two middle quartiles of a Four Equal Groups display was focusing on a part or area of this interval. Importantly, the reading of data stacks in terms of shape was constituted as legitimate despite the differences between Suzanne's and Brad's interpretations on the one hand and Val's on the other hand. It would appear that Val interpreted Suzanne's explanation as a description of a method or a procedure for comparing collections of data points rather than as a way of reasoning about the way the data were distributed.

Later in the discussion, one student also mentioned that he had not compared the datasets in terms of high values because these were the exceptions.[12] The following exchange occurred when the teacher asked whether it was reasonable to compare the two datasets in terms of the high values of corresponding data stacks:

Teacher: What do you think about comparing, looking at these people up here. This person, this person, this person, this person, this person (points to the highest values on the display of data for men), and comparing them to these people over here (points to the highest values on data display for women). Would that give you a pretty good indication of how the men and women's salaries compare, by looking at those?

Kim: No.

Teacher: Why not, Kim?

Kim: Those are more like space out there, like you know what's that man, that Apple man?

Teacher: Bill Gates?

Kim: Yeah,

Suzanne: Ha, that rich dude!

Teacher: I'm afraid he's not even on this. So you're saying those…

Kim: They're just off there like they're just "smarter than the average bear."

Teacher: So you are saying, if he happened to be one of the people, then it would change the whole thing, just one person?

Kim: Yes, that's why I said, don't go by the people on the extremes. Go by the people that are clumped together.

The lack of challenges to or questions about Kim's argument served to legitimize her view that the high values were atypical. Clearly, this exchange contrasts sharply with prior discussions of the relative merits of discerning overall patterns in stacked data by focusing on the medians or on high and low values. As we have noted, most of the students had previously reasoned that a data stack was a collection of data points that occupied a particular interval rather than a distribution that had shape.

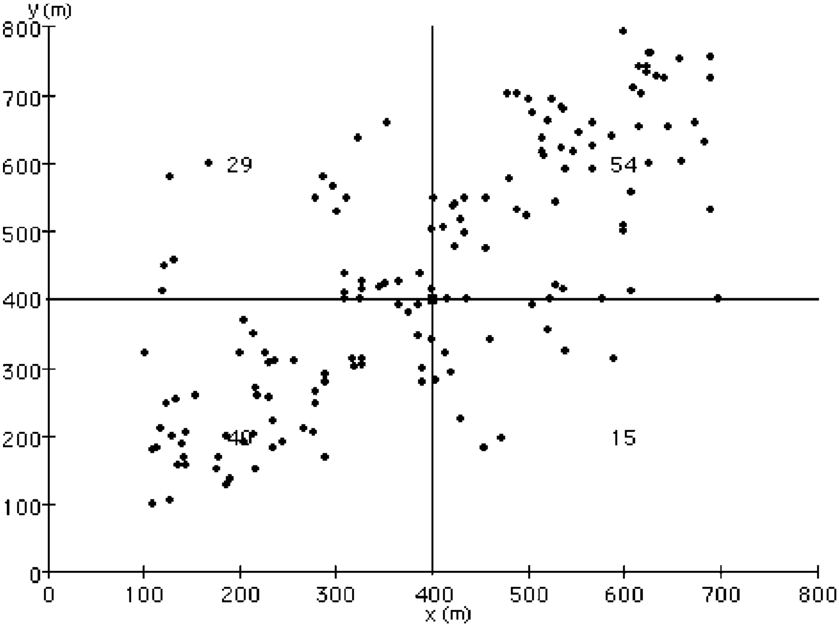
Later in the discussion, the teacher asked the students to describe the overall trends in the two datasets, and it eventually became established that although salary increased with years of education for both men and women, the rate of increase was

[12] Clearly, it can be appropriate when conducting exploratory data analysis to focus on extreme values. In the case of the salary data, for example, the phenomenon of a glass ceiling for women might be investigated by exploring whether there are differences in the salaries of the highest earning men and women who have comparable levels of education. We interpret the student's comment as indicating that he considered it inappropriate to focus on extreme values when identifying overall patterns.

greater for men. It should also be noted that both here and throughout the discussion, the students' contributions were typically grounded in the situation from which the data were generated. For example, they routinely referred to years of education and salary levels when explaining their analyses. In addition, a number of students made comments in which they attempted to account for the differing patterns in men's and women's salaries. This grounding continued when the teacher raised the issue of whether the conclusions that the students had drawn from the data held for the entire country. A number of students raised concerns that related to the size and the representativeness of the samples. For example, one student said that the size of the datasets was too small compared with the number of people in America, and another commented that one dot represented over 1 million people (he later clarified that this was a ratio when challenged by another student). Several students also commented on the need to choose people randomly "to make sure they were not all doctors." Earlier in the discussion, two other students had also made comments that related to the soundness of the design. One had noted that years of work experience could have made a difference, and another had said that although she was not questioning that men were paid more than women, the difference could be because of years on the job. We viewed this range of contributions as encouraging in that they indicate that most of the students had some awareness that the legitimacy of the inferences drawn from data depends crucially on the design of the data creation process.

During the debriefing meeting held immediately after this classroom session, we concluded that our modified conjecture about the reading of the Grids and the Four Equal Groups displays in terms of shape was viable. In addition, we concluded that the revised learning trajectory might also be viable. Ideally, we would have continued to investigate the ways in which the students structured stacked data before exploring the means of supporting a transition to scatter plots in which the data were not stacked. However, this plan was not feasible because only two classroom sessions remained in the time period that we had negotiated with the school district for working with the students. We therefore decided to develop a performance assessment activity that involved comparing two scatter plots. In this activity, the students assessed the relative merits of two speed-reading programs called G1 and G2 by comparing scatter plots of individual pre- and postprogram reading speeds measured in words per minute (see Figure 23).

During the discussion in the next classroom session of the data creation process, the teacher responded to a student's question by clarifying that these were measures of maximum reading speed "with comprehension" before and after enrollment in the programs. When the students then began to work at the computers, all but one pair used the Cross option. In doing so, they left the Cross in the default position as shown in Figure 23 and interpreted the quadrants of each scatter plot in terms of qualitative changes in reading speeds (e.g., improvement, the same, worse). They then compared the reading programs in terms of the number of data points in the corresponding quadrants of the two scatter plots while



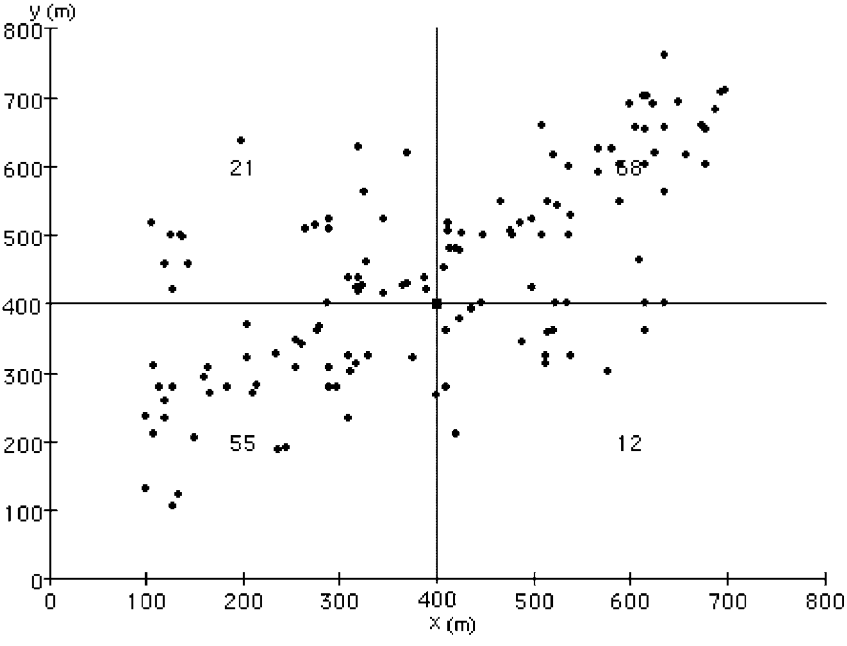


FIGURE 23 Speed reading data with Cross option; G1 (top), G2 (bottom).

adjusting for the unequal size of the datasets. In doing so, they developed descriptions of the effectiveness of the reading programs that were cast in terms of a unidimensional set of qualitative categories. We were initially surprised by these analyses because few of the students had previously used the Cross option and data organized in this way had not previously been the focus of a whole-class discussion. Their use of this option rather than the Grids or the Equal Groups options suggests either that they did not anticipate the possibility of structuring the scatter plots as distributions of univariate distribution or that they did not consider this way of structuring the data to be relevant when assessing the two reading programs. Our choice of data was, in retrospect, unfortunate and indicates that we cannot rule out the latter possibility. We therefore stress that the students were able to develop arguments that they considered both adequate and relevant to the question at hand by using the Cross option. In this regard, their analyses involved a significant advance when compared with those we had observed previously in which they used options such as Four Equal Groups to reduce scatter plots to lines. In contrast to the prior analyses, they treated the two scatter plots as texts about the reading programs.

The one pair of students who analyzed that data in an alternative way used the Four Equal Groups option. When questioned by a researcher, they could explain both why people were in a particular slice (i.e., their preprogram reading speeds were in a particular interval) and what data in different locations within a slice indicated (i.e., some had low reading speeds after the program and some had high reading speeds). They also explained that they had used the Four Equal Groups option so that they could see where "most of the people were." When the researcher asked if they would call this a hill, one of the students said he would not because the data "wasn't on a line." He instead spoke of the majority and explained that this was where the data were bunched up. We interpret his response as indicating that the image of a hill was, for him, specific to stacked data that was piled up "on a line" in a similar manner to univariate data in the axis plot inscription of the second minitool. He saw patterns in the scatter plots according to the "bunched-up-ness" of data (i.e., the relative density of data points) rather than shape.

The first pair of students who presented their analysis during the whole-class discussion explained that they had used the Cross option. In this discussion, it quickly became established that it was illegitimate to compare the total number of data points in corresponding quadrants because the datasets were unequal. The students first explained that the upper left quadrant was "a good quadrant" because the initial reading speed was less than 400 words per min (wpm) and the final reading speed was more than 400 wpm. When another student noted the differing numbers of people in the two programs, the presenting students explained that the G1 program had more people in the good quadrant even though more people had enrolled in G2. In response to a further question, they agreed that it was insufficient to focus only on people who did well and compared the remaining three quadrants in a similar qualitative manner. It was apparent from their explanation that they had taken into account the location of data points within quadrants. For example, they classified the upper left-hand quadrant as “a good quadrant” because the reading speeds of most people in that quadrant of G1 had increased by an amount that they considered significant. However, there was no indication that they viewed the Cross as organizing the scatter plots into two vertical slices. They and the other students who used the Cross option appeared to be comparing univariate datasets (i.e., change in reading speed) rather than bivariate datasets (i.e., relations of covariation).

The next student who presented his solution explained that he had analyzed the data in a similar manner, but had adjusted for the unequal size of the datasets by calculating the percentage of the data in the four quadrants for each dataset. The teacher then asked the students, on the basis of this analysis, how much improvement they could guarantee someone who entered one of the programs with a particular reading speed such as 700 wpm or 300 wpm. In doing so, she attempted to initiate a shift from the interpretation of the Cross as partitioning a scatter plot into four categories to the view that it organized data into two broad slices. She was successful in this regard and, in the ensuing discussion, the student who had calculated percentages acknowledged that he had no idea from his analysis whether the reading speeds of the people whose initial speed was 700 wpm went up or down.

Against this background, the pair of students who had used the Four Equal Groups option presented their analysis by focusing on people who entered the programs reading 300 wpm. In an exchange that involved several other students, they established that 75% of these people’s reading speeds stayed the same or increased for G1, whereas only 50% did so for G2. Importantly, given our instructional agenda, it appeared to be established as normative in this conversation that a slice in the Four Equal Groups display constituted a distribution of postprogram reading speeds for people whose initial reading speeds were in a particular interval.

A final issue addressed in the discussion concerned the usefulness of the Cross, Four Equal Groups, and Grids options. There appeared to be some agreement that the Cross option was useful in giving an overall view, whereas the Four Equal Groups and the Grids options were useful in showing where the data were relatively dense. For example, one student made the following comments about the Grids option:

Teacher: Is this useful at all?

Shane: Yeah, kind of. You can see where the concentrations of people are. And where you are more likely to end up. Depending on where you’re reading. This is almost a combination of the two.

Teacher: Yeah, that’s a nice way to look at it. So, Shane says he can see where the concentrations are and how do you know where the concentrations of people are?

Shane: Where there’s a larger number.

This student’s use of the term “concentrations” appears to be analogous to the way that the students who had used the Four Equal Groups options had spoken of the majority in that both seemed to involve a sense of the relative density of data.

Given the challenging nature of the analysis that we had asked the students to conduct, we viewed this final discussion as again indicating that the revised learning trajectory might be viable. There were, for example, indications that the normative interpretation of data stacks as univariate distributions that have shape might make it possible for students to come to view the slices of a scatter plot as distributions.[13] As we have argued, this is central to treating a scatter plot as a bivariate distribution rather than as merely a dispersed collection of data points. The crucial role of the teacher in supporting and organizing the transition from stacked to unstacked data is also apparent from the discussion. In addition, there was some evidence that this transition might involve a shift from shape, which indicates how the data are distributed, to notions of the majority and concentration, which involve a more direct sense of relative density. However, we did not have an opportunity to investigate this issue fully due to time constraints. For example, it is conceivable that, with appropriate support, some of the students might have interpreted the slices of a scatter plot in terms of shape. This, however, remains an unresolved conjecture.

WHAT WE LEARNED FROM THE DESIGN EXPERIMENTS

In stepping back from the local pedagogical decisions we made to synthesize what we learned, we first return to the discussion of the design experiment methodology. In that discussion, we distinguished between daily minicycles and longer term macrocycles that span an entire design experiment. The account we have given of the actual learning trajectory that was realized in the classroom focuses on the minicycles and the daily testing and revising of conjectures. In the remainder of this article, we take this analysis of the actual learning trajectory as data to complete a macrocycle of design and analysis that spans the entire experiment. Our intent in doing so is to formulate key aspects of a new learning trajectory that could serve as the basis for both a future design experiment and instruction in other classrooms.

In clarifying what we learned from the design experiment, we distinguish between two general aspects of statistical analysis, exploratory data analysis (EDA) and statistical inference.[14] As G. W. Cobb and Moore (1997) noted, EDA puts aside the

[13] For ease of comprehension, we speak of “the normative interpretation of data stacks” as shorthand for “the students’ prior participation in the establishment of the normative interpretation of data stacks.”

[14] We do not focus on the third major aspect of statistical analysis, data creation, in this article because it is dealt with in another article (Tzou, 2000).

question of whether a dataset represents any larger universe and has as its purpose the search for interesting trends and patterns in particular datasets. It should be clear that the search for patterns in bivariate data that were significant or relevant with respect to a question or issue at hand was a major focus of the design experiment. Our intent in designing the third minitool was to make it possible for bivariate datasets to become distributions in which trends and patterns could be discerned.

In contrast to EDA, statistical inference involves drawing conclusions from data that apply to a population. These inferences are probabilistic and ultimately involve the notion of a sampling distribution. We took an initial, tentative step on a trajectory that could eventually lead to this relatively sophisticated notion when we investigated which aspects of datasets the students thought were relatively stable. The issues we addressed with the students were probabilistic in that we asked them to predict how aspects of a dataset might vary if the data creation process were repeated. For ease of explication, we first discuss what we learned about EDA before turning to statistical inference.

Structuring and Organizing Bivariate Data

The insights gleaned from the eighth-grade design experiment have implications for the revision of the prior instructional sequence, which focuses on univariate data, as well as for a learning trajectory that aims at bivariate data, as distributions. We discuss the prior instructional sequence first by considering the starting points for the new learning trajectory and then by outlining three subsequent phases in the potential mathematical development of a classroom community that culminate with bivariate datasets as distributions.

Starting points: univariate datasets as distributions. In preparing for the eighth-grade design experiment, we conjectured that the normative basis for communication that would constitute the starting points might include:

- Using the hill metaphor to describe the types of datasets inscribed as line plots.
- Comparing univariate datasets by structuring them in terms of perceptually based patterns.
- Reasoning multiplicatively about datasets structured in this way in terms of qualitative proportions.

Observations that we made during the first few days of the design experiment when the students used the second minitool indicated that these conjectures were well founded. In addition, the use of Equal Interval Widths and Four Equal Groups displays to compare univariate distributions became normative within the first few classroom sessions. Displays of this type became texts from which the distribution of the data could be read. This advance proved crucial later in the design experiment because it made it possible for stacked data displayed in the third minitool to be constituted as a series of univariate distributions rather than as collections of data points that occupied particular intervals. As we saw, data stacks first came to be treated as distributions in public classroom discourse when the teacher and the students read their shape from the Four Equal Groups and the Grids displays.

The crucial role of shape, which we did not anticipate when preparing for the design experiment, reveals a limitation of our prior work with the students. As we noted, discussions of data organized into Four Equal Groups in the second minitool typically focused on the percentage of the data in particular intervals rather than on shape. Arguments framed in this way can clearly be appropriate when describing differences among univariate datasets. However, discourse of this type provides an inadequate basis for the subsequent emergence of bivariate data as distributions in that it can support a calculational orientation to univariate data rather than a conceptual orientation that is grounded in imagery of how the data are actually distributed.[15] Given our failure to systematically support this conceptual orientation, we were initially surprised when we discovered relatively late in the design experiment that most of the students could easily read the shape of stacks of data from the Grids and the Four Equal Groups displays. The students might have developed this competence when they used the second minitool despite the absence of explicit support. Alternatively, they may have done so as they participated in discussions of the reaction time data. Although some of these data stacks were skewed, they almost invariably had a hill-like shape. The students' familiarity with these datasets might have been such that they developed a feel for how the data were typically distributed and thus the sort of shape the datasets would have if they were inscribed in the second minitool.

These considerations indicate the importance of orienting discussions toward shape for the entire period of time that students analyze univariate data using the second minitool. This proposal in turn requires that thought be given to the characteristics of the datasets that students analyze. We say this because our use of "irregularly shaped" datasets in the latter part of the seventh-grade design experiment may have encouraged talk of percentages of data points in various intervals rather than shape. On reflection, it might have been more productive if we had used datasets whose shapes were smoother in order to support the emergence of symmetric, hill-like datasets (i.e., normally distributed data) as an initial benchmark or point of reference. Other data shapes (e.g., skewed and bimodal data distributions) could then have been contrasted with this reference shape to gradually develop an

[15] This distinction between calculational and conceptual orientations is taken from the work of Thompson et al. (1994), who developed these notions when analyzing mathematics teaching.

interrelated network of paradigmatic data distributions. In such an approach, talking and reasoning about data in terms of quantitative proportions and percentages (i.e., relative frequency) would still be important. However, discourse of this type would emerge as a way of mathematizing shape rather than as an alternative to focusing on shape.

Our second observation that has implications for the starting points of the new learning trajectory concerns the notion of the relative density of univariate data. In addressing this issue, we should clarify that when we speak of shape, we are not referring to a mere figural image of data inscribed as a line plot. Instead, when talk of shape first emerged in the seventh-grade design experiment, it signified a data distribution that was structured multiplicatively and could be reasoned about in terms of qualitative proportions (P. Cobb, 1999; McClain & Cobb, 2001b; McClain et al., 2000). As we have noted, the students typically used the term majority to indicate the proportion of the data in a hill. The teacher and students also used this term frequently in the latter part of the eighth-grade experiment when they discussed both stacked data and scatter plots. However, the normative meaning of the term appeared to have shifted subtly. Whereas majority initially meant an appreciable proportion of a dataset, it was later used in the eighth-grade experiment to indicate the location of a hill where the data were bunched up.

The distinction between these meanings might, at first glance, seem minor in that the observation that a hill that occupies a relatively small part of the range but comprises a relatively large proportion of the data implies that the data are bunched up in that part of the range. However, this implication did not come to the fore in public discourse when students initially spoke of the majority. We therefore conjecture that we might be able to improve the approach we took in the seventh-grade experiment by explicitly attempting to support this shift in normative meaning. This proposal would involve building on discussions of shape to make the relative density of data in various parts of the range an explicit topic of conversation. There is some indication from the eighth-grade experiment that it might be productive to base these discussions on types of data with which the students have become familiar on the basis of first-hand experience, as was the case with reaction time data. The intent in taking such an approach would be to guide the emergence of relative bunched-up-ness or relative density as a characteristic of univariate data distributions that would complement and enrich a focus on shape.

Developing ways of inscribing bivariate data. In building on the starting points we have outlined, there is little indication that we need to modify the first part of the learning trajectory that we formulated at the outset. As we saw, our conjectures about the role of instructional activities in which the students developed and refined inscriptions of bivariate data proved to be viable. For example, the interpretation of bivariate data as consisting of the measures of two attributes of each of a number of cases did become normative. Further, the convention of inscribing such data as dots in a scatter plot emerged relatively easily and enabled us to introduce the third minitool.

Stacked data as bivariate distributions. It is apparent from the analysis we have presented of the design experiment that it was counterproductive to ask the students to analyze scatter plots when the third minitool was first introduced. We did not begin to realize our instructional agenda until we asked the students to analyze stacked data and oriented discussions to the shape of data stacks that were organized using the Grids and the Four Equal Groups options. In the course of these discussions, the stacks were constituted as univariate distributions rather than as collections of data points. The second phase of the new learning trajectory in which students first use the third minitool capitalizes on these observations. Our analysis suggests that the initial instructional activities should involve stacked data with which students have a first-hand familiarity, as was the case with reaction time data in the design experiment. In addition, because the immediate goal is to guide the constitution of stacks as distributions, it might be worthwhile to limit the number of stacks to three or four (e.g., reaction time data for ages 20, 40, and 60 years), and to frame the instructional activity so that the purpose is to use the third minitool to compare stacks rather than to search for trends and patterns across the stacks. We conjecture that this might facilitate a focus on shape and relative density. Follow-up analyses might then be concerned with the covariation of the two sets of measures (i.e., trends and patterns in the distribution of the univariate data distributions). This, we conjecture, might support the constitution of the entire dataset as a bivariate distribution. To the extent that this occurs, later instructional activities might involve a greater number of data stacks and might focus on patterns across stacks from the outset (e.g., data, measured monthly, on the resting heart rates of a group of individuals who have been enrolled in an exercise program that is to last 8 months).

It should be clear that shape and relative density are central to this phase of the new learning trajectory. In this regard, it is worth noting that because statistical covariation involves coordinating the variation of two sets of measures, it is often viewed as being two-dimensional and thus as being relatively transparent in scatter plots. However, the analysis we have presented leads us to argue that proficient statistical analysts' imagery of covariation is, metaphorically speaking, no more two-dimensional than their imagery of univariate distributions is one-dimensional. This is clearer in the case of univariate data in that inscriptions such as stem plots and line plots involve, for the proficient user, a second dimension, which indicates relative frequency. In the case of bivariate data, however, scatter plots do not provide such direct perceptual support for a third dimension corresponding to relative frequency.[16] Instead, it appears that proficient analysts read this third dimension from

[16] This notion of an implicit third dimension in bivariate data was first brought to our attention by Patrick Thompson (personal communication, August 22, 1998).

the relative density of the data points. The difficulties we encountered in the design experiment indicate that it is unreasonable to expect students to read scatter plots in this sophisticated manner from the outset. We saw, for example, that when they first used the third minitool, they typically reduced scatter plots to lines that signified fixed relationships of covariation rather than conjectured relationships about which the data were distributed. In contrast, they and the teacher literally introduced a third dimension when they later discussed the shape of data stacks in two-dimensional data displays. It was apparent that the shape of stacks read from the Grids and the Four Equal Groups displays was relatively concrete for most of the students. Further, it quickly became normative that the shape of a stack could itself be read in terms of the relative density or frequency of the data. The intent of the second phase of the new trajectory is to support the emergence of this way of reasoning about stacked data.

Scatter plots as bivariate distributions. The final phase of the new trajectory is concerned with the structuring of scatter plots (i.e., unstacked data). Our analysis of the last few sessions of the design experiment indicates that a continued focus on the relative density and perhaps the shape of data might be crucial in supporting the transition from stacked to unstacked data. This in turn suggests that it could be important to orient initial discussions in this phase of the trajectory toward the distribution of data within slices of a scatter plot organized in terms of the Grids or the Four Equal Groups option. An initial analysis might involve data on, say, the length of time that a group of people spend brushing their teeth each day and the amount of plaque on their teeth.[17] We saw from the design experiment that, in such a case, it might be productive to raise issues that lead to a focus on the people who brush their teeth for a certain length of time (i.e., a data slice). Further, it appears important to discuss both who is in a particular slice (e.g., the people who brush for about 2 min) and what their location within a slice indicates (i.e., the amount of plaque on their teeth). Against this background, questions that involve the comparison of slices might lead to discussions in which the focus is on the distribution of data within slices. This final phase of the new trajectory rests on the conjecture that a concern for the relative density of data within slices might emerge and become normative in the course of these discussions.

In this regard, it is worth recalling that, in the design experiment, we did not have an opportunity to explore whether some of the students might have found it reasonable to talk of the data within a slice as having shape. It might be productive to explore this possibility because shape appeared to have become a relatively

[17] This instructional activity was developed by Cliff Konold and was originally proposed as an assessment task (personal communication, March 9, 1999).

concrete notion for most of the students, indicating that it might support a focus on the relative density of data.

As a further conjecture, we infer from the design experiment analysis that this initial emphasis on the distribution of data within slices might provide a basis for a subsequent focus on trends and patterns in an entire dataset. We can clarify the normative ways of talking and reasoning about data that we hope will emerge by employing the metaphor of shape. Suppose, in particular, that the data within slices in the plaque example have hill-like shapes. In terms of the metaphor, a ridge might be seen running across the dataset where the data within slices are relatively dense. Further, the ridge might be viewed as indicating a conjectured relationship of covariation between brushing time and amount of plaque about which the data are distributed. Imagery of this sort, whether involving shape or a more direct sense of relative density, would appear to support discussions of the strength as well as the direction of the relationship. In addition, this imagery could also support an investigation of cases that do not fit the overall pattern. Further, because data would be viewed as distributed about the ridge vertically rather than diagonally, such imagery would appear to provide a basis for later, more advanced, analyses of statistical covariation that are concerned with finding the line of best fit by minimizing the squares of the deviations of the *y* measures.

As a final observation, we argued when discussing the starting points for the new learning trajectory that it might be important to support the emergence of an interrelated network of paradigmatic univariate data shapes. We speculate that a similar emphasis might also be appropriate for bivariate data. The hypothetical data shape that we described when elaborating the plaque example could constitute an initial point of reference against which other bivariate data distributions could be contrasted. This suggestion indicates both the potential value of instructional activities that involve comparing two scatter plots and the level of planning required when selecting datasets that students might analyze.

In concluding this discussion of the newly formulated learning trajectory, it is worth highlighting our use of the relatively concrete, physical metaphor of shape to talk about bivariate data distributions. We used this metaphor in an attempt to characterize the ways of talking and reasoning about data that we hope might become normative. The metaphor is useful even if actual classroom discourse is more directly concerned with relative density than with shape, in that it emphasizes that we want trends and patterns in bivariate data to become almost tangible for students. Greeno (1991) used the metaphor of a mathematical environment in which tools and resources are readily available for characterizing number sense in particular and mathematical knowing more generally. Sfard (2000b) described mathematical discourse as a virtual reality discourse, to highlight the parallels between this discourse and the ways in which we talk about physical reality. In Greeno's (1991) and Sfard's (2000b) terms, we want students to come to act in a statistical environment or reality in which bivariate data have substance and structure that can be investigated and talked about. The new learning trajectory we have outlined indicates our current conjectures about the means of supporting the gradual emergence of such an environment.

Initial Steps Toward Statistical Inference

The new learning trajectory that we have outlined is concerned with EDA in that it leaves aside the question of whether a dataset is representative of a larger universe and focuses on the search for trends and patterns in particular datasets. Issues relating to statistical inference came to the fore in the design experiment when the teacher and the students discussed the relative stability of the median and the extreme values of univariate data stacks. Our intent in asking the students to predict how these aspects of a dataset might vary if the data creation process were repeated was to support a focus on the center rather than on the extremes of data slices when tracing relationships of covariation in bivariate data. In retrospect, it is clear that this approach was flawed in that it confounded the structuring of particular datasets with relative stability across samples, and thus EDA with statistical inference. Despite this significant limitation, this phase of the design experiment did give rise to insights that can inform further revisions of the learning trajectory.

During the analysis of the design experiment, we differentiated between two distinct orientations to the relationship between data and the median. One of these orientations derived from school-taught methods for finding the median, whereas the other involved the notion of shape and was grounded in the students' prior participation in the seventh-grade design experiment. As we have seen, the questions we posed about the relative stability of the median and the extreme values appeared to support the first of these orientations. It was evident that the students all knew how to find the median of a dataset. In addition, they viewed the computer as carrying out this calculational process for them when they used the Two Equal Groups and the Four Equal Groups options. However, there was every indication that the result of this process was, for all but one of the students, a value in the interval occupied by a dataset that was at the same level as individual data points. In contrast, the one remaining student developed arguments in which he reasoned about the relation between changes in individual data points and changes in the median. From this we inferred that the median was, for him, a structural feature of a dataset that depended on all the data values.

We documented that all but this student continued to view the median as point in the interval occupied by a dataset when the students participated in a series of discussions that focused on the relation between data and the median. We mention that in addition to the instructional activities we described when reporting these discussions, the students also engaged in a simulated sampling activity. In this activity, they each generated samples by randomly selecting 9 reaction time measures from a set of 100 such measures. The students were surprised that the medians of their samples were relatively close. However, for most of the students, the activity appeared to be little more than an empirical demonstration that the median was more stable than the extreme values. In contrast, the student who had previously viewed the median in structural terms went on to find, on his own initiative, the highest and lowest possible values of the median of samples of 9 measures (i.e., the range of the sampling distribution for the median). In doing so, he again reasoned about the relation between changes in an entire sample of data values and changes in the median. The continued epistemological gulf between this student and the others indicates the inadequacy of the instructional approach that we took.

With hindsight, we have come to view the general orientation to data and the median that we unwittingly encouraged as the primary source of the impasse that the students and we arrived at midway through the design experiment. Thus, although the instructional activities used in this phase of the experiment were clearly not exemplary, we contend that a preoccupation with the limitations of specific activities misses the larger issue. This claim is supported by observations that we made once the teacher had guided the emergence of a normative orientation to shape so that data stacks became constituted as distributions rather than collections of data points. As we saw, the expectation that the shape of a data stack would be relatively stable if the process of generating the data were repeated quickly became established as normative. An exchange that we observed earlier goes some way toward explaining this development. This exchange occurred shortly after the students had each generated 10 measures of their individual reaction times and the teacher had drawn plots of 10 of the students' measures (see Figure 16). As we reported, the teacher asked the students if they could use these data to predict the reaction time of another eighth grader. The students all appeared to reject this possibility, arguing that their reaction times were different and that it would depend on the particular eighth grader. In contrast, when the teacher next asked the students if they could predict the reaction times of another group of 10 eighth graders, most indicated almost immediately that this was feasible. As one student noted by way of justification, "Some will be slow, some will be extremely fast like, you know, like some of us." This argument, which was constituted as legitimate in the classroom, appeared to involve a concern for variability if not for the way that the data were distributed. We speculate that it was this expectation of similar variability across samples that underpinned the later prediction that the shape of a data stack was relatively stable.

These were indications that this expectation about the stability of shape can support the realization that the median is also relatively stable. In the case of the reaction time data, for example, the interpretation of the median in a Four Equal Groups display as indicating the location of a hill in the data became established as normative with little difficulty. The median could therefore be viewed as relatively stable because it was located in a particular position within a stable data shape. It is important to stress that this insight does not necessitate a structural interpretation of the relation between changes in the data and the changes in the median. Instead, it appeared that, for most of the students, the median was a location within a data shape rather than a structural characteristic of an entire dataset.

The reflections we have presented lead us to question the value of initially approaching statistical inference by attempting to build on school-taught methods for finding the median (or the mean for that matter). Instead, it appears that the shape of univariate data distributions might constitute a more promising starting point. The support it provides in terms of concrete imagery can be contrasted with the calculational orientation fostered by school-taught methods. This observation has implications for the revision of our prior work with the students as well as for the new learning trajectory. We have already argued that shape should feature prominently in students' analyses of univariate data. Once an orientation to the shape of data has become normative, it might be worthwhile to investigate with students the stability of shape if the data creation process is repeated. This in turn could lead to discussions of the relation between the shape of a data sample and that of data for the population. In the course of these discussions, it might also be possible to consider the relation between the median of a sample (i.e., a sample statistic) and that of the population (i.e., a population parameter), provided that shape serves to anchor the conversation. Once established, discussions of this type could be continued as a matter of course when students later analyze bivariate data. We stress that we envision these conversations as being conducted in relatively informal terms. We speculated that the orientation to sample–population relations engendered by such exchanges might provide a starting point for a subsequent learning trajectory that aims squarely at the challenging notion of sampling distribution.

CONCLUSIONS

The potential pedagogical significance of the issues we have discussed in this article is not restricted to the teaching of statistics at the middle-school level, given the relatively sophisticated ways in which the students reasoned about data in the latter sessions of the design experiment. The National Council of Teachers of Mathematics (2000) recommended, for example, that analyses of bivariate data that focus on statistical covariation should be delayed until the high school level. As a further contribution, this article also goes some way to addressing Shaughnessey, Garfield, and Greer's (1997) call for longitudinal case studies of statistics learning and teaching, thereby complementing previous investigations of this type (Biehler & Steinbring, 1991; de Lange et al., 1993; Hancock, Kaput, & Goldsmith, 1992; Lehrer &

Romberg, 1996). As Shaughnessey et al. (1997) noted, such studies can make important contributions to the development of a theoretical framework for statistics instruction. It should be clear that, in the analysis we presented, theoretical considerations came to the fore when we developed the rationales of successive revisions of the learning trajectories. Convergence with a second of Shaughnessey et al.'s recommendations becomes apparent when we focus on the revised learning trajectory that we have outlined. In concert with their arguments, we suggested that probability should be included in what they termed data handling (i.e., EDA). As they noted, questions that involve resampling give rise to issues that are probabilistic in nature.

The revised, yet still hypothetical, learning trajectory that we outlined is also consistent with G. W. Cobb and Moore's (1997) discussion of the order in which the three major aspects of statistical analysis should be addressed. They recommended beginning with methods for exploring and describing data (i.e., EDA) before focusing on data creation and then finally moving to statistical inference. Although we found it essential for the teacher to talk through the data creation process with students from the outset, we also clarified that the initial instructional focus in the revised trajectory is on ways of structuring and organizing univariate data. We found that issues relating to the data creation process, such as methods of sampling and controlling extraneous variables, gradually emerged as topics of conversation in a relatively natural way (Tzou, 2000). These conversations about data creation indicate that it became normative that conclusions could not be drawn from data unless their generation was sound. Lehrer and Romberg (1996) spoke of the reflexivity of data creation and data analysis to highlight this interdependence.

Consistent with G. W. Cobb and Moore's (1997) recommendations, issues such as sample–population relations that underpin statistical inference are delayed in the final trajectory until univariate datasets have been constituted as distributions that have shape. It is important to acknowledge that G. W. Cobb and Moore's proposals dealt with the teaching and learning of statistics at the college level. As a consequence, they were concerned with more formal aspects of statistical inference such as sampling distributions, confidence intervals, and significance tests. We, in contrast, discussed a far more informal approach to statistical inference. Despite this difference, we find it encouraging that the broad outline of G. W. Cobb and Moore's proposal appears to be relevant to statistical analysis at the middle-school level.

Turning now to consider methodological issues, the account we gave of the eighth-grade design experiment serves to illustrate a way of developing and improving instructional designs when the research base is relatively limited.[18] The initial learning trajectory that we formulated when preparing for the design experiment was highly provisional and a number of our initial conjectures proved

[18] McGatha (2000) addressed this issue in considerably more detail by analyzing the research team's learning as its members prepared for and conducted the seventh-grade design experiment.

to be unviable. However, the ongoing process of testing and revising conjectures enabled us eventually to formulate a new trajectory that was empirically grounded in our work in the classroom (Gravemeijer, 1998). The key point to note is that, in this bootstrapping approach, we did not wait for the development of an adequate research base, but instead outlined and began testing an initial instructional design. We contend that what we learned about the learning and teaching of statistics while doing so contributes to a nascent, but gradually emerging local instructional theory that can inform both future investigations and instructional practice in other classrooms. This approach, in which theoretical analyses and instructional designs coemerge, is captured by the adage that if you want to understand something (e.g., students' statistical reasoning), try to change it, and if you want to change something, try to understand it (De Corte, Greer, & Verschaffel, 1996).

As a related observation, it is worth clarifying that the account we gave of the students' reasoning is cast in situated terms. By this, we do not mean that the meanings the students developed in the course of the design experiment had no currency beyond the classroom. Instead, we contend that the students' mathematical learning was situated with respect to the classroom microculture that they and the teacher constituted in the course of their ongoing interactions. We document this evolving microculture by first reporting the social and sociomathematical norms that were established in the classroom and then tracing the normative mathematical meanings that emerged over the course of the experiment. The students' participation in both the regeneration and the evolution of these norms and meanings constituted the immediate social situation of their learning. It should be clear that in adopting this interpretive stance, we did not view particular students' reasoning as inherent characteristics or properties of them as individuals. As a consequence, we did not interpret instances of students failing to learn as expected as indicators of their conceptual inadequacies. Instead, we interpreted these instances as indicating inadequacies in our current instructional design. Thus, we did not interpret most of the students' failure to develop a structural view of the relation between data and the median as indicating that their understandings of data were too immature for them to benefit from the instructional activities we had developed. Instead, we concluded that the instructional approach we had taken was deeply flawed in that it attempted to build on school-taught methods.

We argue that this perspective on students' learning is a strength rather than a weakness in that we viewed their reasoning as situated with respect to the type of discourse, tools, and resources that served as means of supporting its development. This made it possible for us to develop new design conjectures by drawing on our ongoing analysis of classroom events. As a consequence, a difficulty that typically arises when students' reasoning is analyzed exclusively in individualistic psychological terms failed to materialize: the difficulty of figuring out what the analysis might mean for design and instruction. Given our interest in formulating and improving instructional designs, we find this characteristic of a situated perspective extremely attractive.

As a final observation, we conjecture that in addition to informing the revision of instructional designs and providing guidance for instructional practice in other classrooms, analyses of the type that we illustrated can also serve as important means of supporting the development of professional teaching communities (Ball & Cohen, 1996; Hiebert & Wearne, 1992).[19] As we saw, the analyses serve to justify the instructional sequences that are developed in the course of design experiments in terms of (a) the trajectory of the classroom community's mathematical learning and (b) the means of supporting that learning. If the sequences were justified solely with traditional experimental data, teachers would know only that the sequences had proved effective elsewhere, but they would not have an understanding of the underlying rationale that would enable them to adapt the sequences to their own instructional settings. In contrast, the type of justification that we favor offers the possibility that teachers will be able to adapt, test, and modify the sequences in their classrooms. In doing so, they can then contribute to both the improvement of the sequences and the development of local instructional theories rather than merely being the consumers of instructional innovations developed by others. As part of our current work, we are investigating the feasibility of supporting the development of professional teaching communities by collaborating with two groups of teachers who work in urban school districts. In doing so, one of our primary goals is to ensure that the teachers view implementation as a process of conjecture-driven adaptation in which they test and refine pedagogical approaches that have proven effective elsewhere.

In conclusion, we extend this view of implementation as conjecture-driven adaptation to our fellow researchers by stressing that both the design experiment methodology we illustrated and the specific learning trajectory we have described are conceptual tools that were developed while addressing specific problems and issues. We therefore assume that others will adapt these tools as they apply them to their own problems of interest, and in the process, contribute to the ongoing refinement and improvement of the tools.

[19] It could be argued that the forms of instruction developed in the course of a design experiment are unfeasible for any teacher working alone. We acknowledge, for example, that the entire research team in effect constitutes a collective teacher with some members of the team actually teaching while others observe and analyze classroom events. The demands of this collective activity are, however, balanced by the possibility that the collaborating teachers will be able to capitalize on our learning as represented by instructional sequences and learning trajectories. This conjecture about the proposed role of instructional sequences as a means of supporting the development of professional teaching communities is discussed in some detail by P. Cobb and McClain (2001).

ACKNOWLEDGMENTS

The analysis reported in this article was supported by the National Science Foundation under Grant No. REC 9814898 and by the Office of Educational Research and Improvement (OERI) under grant number R305A60007. The opinions expressed do not necessarily reflect the views of either the Foundation or OERI.

REFERENCES

Atkinson, P., Delamont, S., & Hammersley, M. (1988). Qualitative research traditions: A British response to Jacob. *Review of Educational Research, 58,* 231–250.

Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher, 25*(9), 6–8, 14.

Biehler, R. (1993). Software tools and mathematics education: The case of statistics. In C. Keitel & K. Ruthven (Eds.), *Learning from computers: Mathematics education and technology* (pp. 68–100). Berlin: Springer.

Biehler, R., & Steinbring, H. (1991). Entdeckende Statistik, Stenget-und-Blatter, Boxplots: Konzepte, Begrundungen and Enfahrungen eines Unterrichtsversuches [Explorations in statistics, stem-and-leaf, boxplots: Concepts, justifications, and experience in a teaching experiment].*Mathematikunterricht, 37*(6), 5–32.

Bowers, J., Cobb, P., & McClain, K. (1999). The evolution of mathematical practices: A case study. *Cognition and Instruction, 17,* 25–64.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classrooms. *Journal of the Learning Sciences, 2,* 141–178

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly, 104,* 801–823.

Cobb, P. (1995). Mathematics learning and small group interactions: Four case studies. In P. Cobb & H. Bauersfeld (Eds.), *Emergence of mathematical meaning: Interaction in classroom cultures* (pp. 25–129). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cobb, P. (1999). Individual and collective mathematical learning: The case of statistical data analysis. *Mathematical Thinking and Learning*, 1999, 5–44.

Cobb, P. (2000). Conducting classroom teaching experiments in collaboration with teachers. In A. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 307–334). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cobb, P. (2001). Supporting the improvement of learning and teaching in social and institutional context. In S. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 455–478). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cobb, P., & McClain, K. (2001). An approach for supporting teachers' learning in social context. In F.-L. Lin & T. Cooney (Eds.), *Making sense of mathematics teacher education* (pp. 207–232). Dordrecht, the Netherlands: Kluwer.

Cobb, P., & Whitenack, J. W. (1996) A method for conducting longitudinal analyses of classroom videorecordings and transcripts. *Educational Studies in Mathematics, 30,* 458–477.

Cobb, P., Stephan, M., McClain, K., & Gravemeijer, K. (2001). Participating in classroom mathematical practices. *Journal of the Learning Sciences, 10*(1&2), 113–164.

Collins, A. (1999). The changing infrastructure of educational research. In E. C. Langemann & L. S. Shulman (Eds.), *Issues in education research*. San Francisco: Jossey Bass.

Confrey, J., & Lachance, A. (2000). A research design model for conjecture-driven teaching experiments. In A. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 231–266). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

De Corte, E., Greer, B., & Verschaffel, L. (1996). Mathematics learning and teaching. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 491–540). New York: Macmillan.

de Lange, J., van Reeuwijk, M., Burrill, G., & Romberg, T. (1993). *Learning and testing mathematics in context. The case: Data visualization*. Madison: University of Wisconsin, National Center for Research in Mathematical Sciences Education.

Dörfler, W. (1993). Computer use and views of the mind. In C. Keitel & K. Ruthven (Eds.), *Learning from computers: Mathematics education and technology* (pp. 159–186). Berlin: Springer-Verlag.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *The handbook of research on teaching* (3rd ed., pp. 119–161). New York: Macmillan.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. New York: Aldine.

Gravemeijer, K. E. P. (1994). *Developing realistic mathematics education*. Utrecht, The Netherlands: CD-B Press.

Gravemeijer, K. (1998, April). *Developmental research: Fostering a dialectic relation between theory and practice*. Paper presented at the research presession of the annual meeting of the National Council of Teachers of Mathematics, Washington, DC.

Greeno, J. G. (1991). Number sense as situated knowing in a conceptual domain. *Journal for Research in Mathematics Education, 22,* 170–218.

Hancock, C., Kaput, J. J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist, 27,* 337–364.

Hiebert, J., & Wearne, D. (1992). Instructional tasks, classroom discourse, and students' learning in second grade arithmetic. *American Educational Research Journal, 30,* 393–425.

Hershkowitz, R., & Schwartz, B. (1999). The emergent perspective in rich learning environments: Some roles of tools and activities in the construction of sociomathematical norms. *Educational Studies in Mathematics, 39,* 149–166.

Hodge, L. L. (2001). *Students' emerging identities as doers of mathematics in two contrasting classroom microcultures*. Unpublished manuscript.

Kaput, J. J. (1994). The representational roles of technology in connecting mathematics with authentic experience. In R. Biehler, R. W. Scholz, R. Strasser, & B. Winkelmann (Eds.), *Didactics of mathematics as a scientific discipline* (pp. 379–397). Dordrecht, The Netherlands: Kluwer.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 6,* 59–98.

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal, 27*(1), 29–63.

Latour, B. (1987). *Science in action*. Cambridge, MA: Harvard University Press.

Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life*. New York: Cambridge University Press.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction, 14,* 69–108.

McClain, K. & Cobb, P. (2001a). The development of sociomathematical norms in one first-grade classroom. *Journal for Research in Mathematics Education, 32,* 234–266.

McClain, K., & Cobb, P. (2001b). Supporting students' ability to reason about data. *Educational Studies in Mathematics, 45*(1–3), 103–129.

McClain, K., Cobb, P., & Gravemeijer, K. (2000). Supporting students' ways of reasoning about data. In M. Burke (Ed.), *Learning mathematics for a new century* (pp. 174–187). Reston, VA: National Council of Teachers of Mathematics.

McGatha, M. (2000). *Instructional design in the context of classroom-based research: Documenting the learning of a research team as it engaged in a mathematics design experiment*. Unpublished dissertation, Vanderbilt University, Nashville, TN.

McGatha, M., Cobb, P., & McClain K. (1999, April). *An analysis of student's initial statistical understandings*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Meira, L. (1998). Making sense of instructional devices: The emergence of transparency in mathematical activity. *Journal for Research in Mathematics Education, 29,* 121–142.

Much, N. C., & Shweder, R. A. (1978). Speaking of rules: The analysis of culture in breach. *New Directions for Child Development, 2,* 19–39.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Nemirovsky, R., & Monk, S. (2000). "If you look at it the other way…" An exploration into the nature of symbolizing. In P. Cobb, E. Yackel, & K. McClain (Eds.), *Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design* (pp. 177–221). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Noss, R., Pozzi, S., & Hoyles, C. (1999). Touching epistemologies: Statistics in practice. *Educational Studies in Mathematics, 40,* 25–51.

Pea, R. D. (1993). Practices of distributed intelligence and designs for education. In G. Solomon (Ed.), *Distributed cognitions* (pp. 47–87). New York: Cambridge University Press.

Roth, W.-M. (1997.) Where is the context in contextual word problems? Mathematical practices and products in grade 8 students' answers to story problems. *Cognition and Instruction, 14,* 487–527.

Roth, W.-M., & McGinn, M. K. (1998). Inscriptions: Towards a theory of representing as social practice. *Review of Educational Research, 68,* 35–59.

Schutz, A. (1962). *The problem of social reality*. The Hague, Holland: Martinus Nijhoff.

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22,* 1–36.

Sfard, A. (2000a). On reform movement and the limits of mathematical discourse. *Mathematical Thinking and Learning, 2,* 157–189.

Sfard, A. (2000b). Symbolizing mathematical reality into being. In P. Cobb, E. Yackel, & K. McClain (Eds.), *Symbolizing, communicating, and mathematizing: Perspectives on discourse, tools, and instructional design* (pp. 37–98). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Shaughnessey, J. N., Garfield, J., & Greer, B. (1997). Data handling. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (Part 1, pp. 205–237). Dordrecht, The Netherlands: Kluwer.

Simon, M. A. (1995.) Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26,* 114–145.

Simon, M. A. (2000). Research on mathematics teacher development: The teacher development experiment. In A. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 335–359). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Simon, M. A., & Blume, G. W. (1996). Justification in the mathematics classroom: A study of prospective elementary teachers. *Journal of Mathematical Behavior, 15,* 3–31.

Steffe, L. P., & Thompson, P. W. (2000). Teaching experiment methodology: Underlying principles and essential elements. In A. Kelly & R. Lesh (Eds.), *Handbook of research design in mathematics and science education* (pp. 267–306). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap*. New York: Free Press.

Suchman, L. A., & Trigg, R. H. (1993.) Artificial intelligence as craftwork. In S. Chaiklin & J. Lave (Eds.), *Understanding practice: Perspectives on activity and context* (pp. 144–178). New York: Cambridge University Press.

Suter, L. E., & Frechtling, J. (2000). *Guiding principles for mathematics and science education research methods: Report of a workshop*. Washington, DC: National Science Foundation.

Taylor, S. J., & Bogdan, R. (1984). *Introduction to qualitative research methods* (2nd ed.). New York: Wiley.

Thompson, A. G., Philipp, R. A., Thompson, P. W., & Boyd, B. A. (1994). Calculational and conceptual orientations in teaching mathematics. In A. Coxford (Ed.), *1994 Yearbook of the National Council of Teachers of Mathematics* (pp. 79–92). Reston, VA: National Council of Teachers of Mathematics.

Tzou, C. (2000). *Learning about data creation*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Voigt, J. (1985). Patterns and routines in classroom interaction. *Recherches en Didactique des Mathematiques, 6,* 69–118.

Voigt, J. (1995). Thematic patterns of interaction and sociomathematical norms. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning: Interaction in classroom cultures* (pp. 163–202). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Wenger, E. (1998). *Communities of practice*. New York: Cambridge University Press.

Wilensky, U. (1997.) What is normal anyway? Therapy for epistemological anxiety. *Educational Studies on Mathematics, 33,* 171–202

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentations and autonomy in mathematics. *Journal for Research in Mathematics Education, 27,* 458–477.