MODELING AS A CORE COMPONENT OF STRUCTURING DATA

CLIFFORD KONOLD
University of Massachusetts Amherst
konold@srri.umass.edu

WILLIAM FINZER
Concord Consortium
wfinzer@concord.org

KOSOOM KREETONG
University of Massachusetts Amherst
kkreeton@educ.umass.edu

ABSTRACT

We gave participants diagrams of traffic on two roads with information about eight attributes, including the type of each vehicle, its speed, direction and the width of the road. Their task was to record and organize the data to assist city planners in its analysis. Successfully encoding the information required the creation of a case, a physical record of one repetition of a repeatable observational process. We analyzed data sheets participants created including the methods they used to bind information together into cases. Overall, 79% of their data sheets successfully encoded the data. Even 62% of the middle school students were able to create a bound structure that could hold the critical information from the diagrams. A majority of these structures involved a hierarchy of cases rather than the "flat" case-by-attribute structure that virtually all statistical software require. Our sense is that participants strove to create a structure that modeled the real-world as closely as they could, constructing cases that corresponded to the different sorts of objects they perceived—vehicles with their characteristics nested within road segments with their characteristics.

Keywords: *Statistics education research; Data modeling; Data organization; Data collection; Tables*

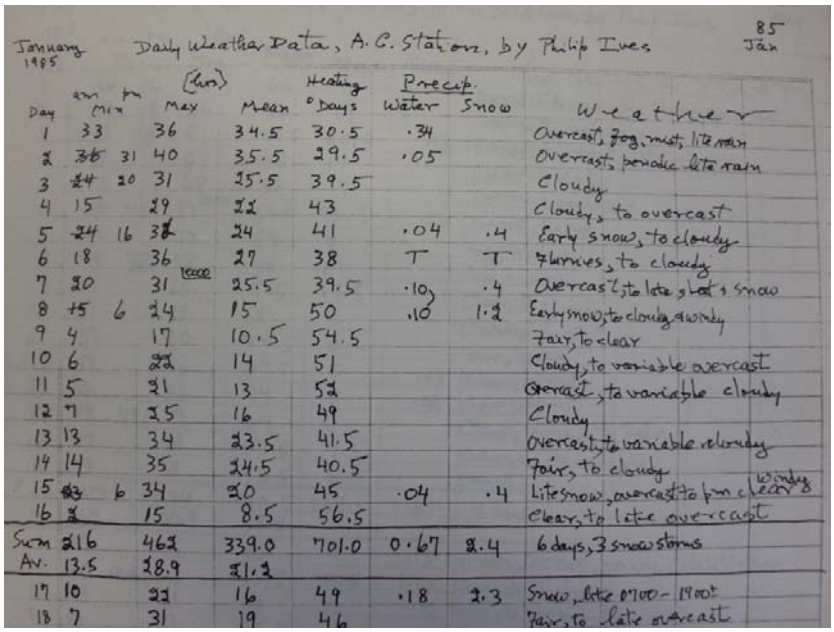
1. INTRODUCTION

1.1. USING CASE TABLES TO STORE AND ORGANIZE DATA

For 44 years, Philip Ives, a biology professor at Amherst College, recorded daily climatological observations from a small weather station just outside his campus office. Every other week he would write up a short summary that would appear in the local paper under his byline. Figure 1 is part of a page from his handwritten notebook in which he logged daily values from January 1985.

There is a lot going on in this table. The body consists of rows containing information about the weather for individual days. Each row is divided into cells, one per column. The columns are labeled with a brief description of the information in that column. The column

'Min' (Minimum Temperature) is divided into two sub-columns, 'am' and 'pm.' If the pm value was lower than the am value, the am value was crossed out and the pm value recorded. Two columns—'Water' and 'Snow'—are grouped together under 'Precip.' The values in two columns—'Mean' and 'Heating °Days'—are computed from values in other columns. The right-hand column consists of notes about the observable weather conditions. The two rows below Day 16 contain summary information for the first half of the month; these were the values included in his biweekly newspaper summary. A title appears at the top, which includes the name of the person making the observations. The month and year are recorded at the top of the page.



January 1905 — Daily Weather Data, A. C. Station, by Philip Ives — 85 Jan

Day	Min am	Min pm	Max	Mean	Heating °Days	Precip. Water	Precip. Snow	Weather
1	33		36	34.5	30.5	.34		Overcast, fog, mist, lite rain
2	~~36~~	31	40	35.5	29.5	.05		Overcast, periodic lite rain
3	~~24~~	20	31	25.5	39.5			Cloudy
4	15		29	22	43			Cloudy, to overcast
5	~~24~~	16	~~38~~	24	41	.04	.4	Early snow, to cloudy
6	18		36	27	38	T	T	flurries, to cloudy
7	20		31	25.5	39.5	.10	.4	Overcast, to lite, slat, snow
8	15	6	24	15	50	.10	1.2	Early snow, to cloudy, windy
9	4		17	10.5	54.5			Fair, to clear
10	6		22	14	51			Cloudy, to variable overcast
11	5		21	13	52			Overcast, to variable cloudy
12	7		25	16	49			Cloudy
13	13		34	23.5	41.5			Overcast, to variable cloudy
14	14		35	24.5	40.5			Fair, to cloudy
15	~~33~~	6	34	20	45	.04	.4	Lite snow, overcast to pm clear (winter)
16	2		15	8.5	56.5			Clear, to late overcast
Sum	216		462	339.0	701.0	0.67	8.4	6 days, 3 snow storms
Av.	13.5		28.9	21.2				
17	10		22	16	49	.18	2.3	Snow, late 0700 – 1900
18	7		31	19	46			Fair, to late overcast

Figure 1. Local climatological observations from the notebook of Philip Ives

Records like these form the bedrock upon which statistical analyses are built. Like most foundations, they typically remain invisible. Even in comprehensive depictions of statistical investigations (Friel & Bright, 1998; Wild & Pfannkuch, 1999), the components involved in making such recordings are generally subsumed under a more general descriptor such as "collect the data." If you inspect curriculum materials for explicit instruction on recording and organizing data or search the literature for research on how students go about or learn these skills, you will find virtually nothing.

Data are most typically recorded in table format. Research on the use of tables in statistical practice has, too, been relatively neglected, perhaps because tables are so ubiquitous that we tend to take them for granted. Indeed, they have been with us a long time. Figure 2 is a table (from Robson, 2003) c. 2028 BCE made on a clay tablet found in Ur, a city located in what today is southern Iraq. The right column contains the labels "lambs," "first-rate sheep," and "billy goats." The column headings are in the fifth row and appear to be personal names. The top three rows are numerals, and the fourth row contains column totals. Thus the table seems to be an account of the number of each type of animal owned by five people. The date of the record appears on the reverse side of the tablet.

Figure 2. Clay tablet from Ur, earliest known use of a table

According to Robson (2003), the invention of the sexagesimal place value system was a precursor to the recording of such information in tables as opposed to lists or narratives, because it allowed for the separation of quantitative and qualitative information. Separating this information and recording them in tables provided "great efficiency both in recording and retrieving structured information" (p. 14). It also facilitated error checking and computation, as one could now quickly add values contained in the same column.

We make a distinction between two types of statistical tables. The type of table that Ives made to record his weather observations we refer to as a "case-data table." Such tables are created primarily to collect and/or hold the raw data and thus are typically not structured in ways that help reveal patterns or trends in those data. The second type of table we will call "summary tables." These typically contain only some of the information collected and are organized to facilitate comparing groups or detecting trends. A number of researchers have explored the use of statistical summary tables (Estrella, 2014; Martí, Garcia-Mila, Gabucio, & Konstantinidou, 2011; Prodromou, 2015) and developed criteria for producing easy-to-read displays (Ehrenberg, 1977). In this article, we explore the ways in which students record and organize data, and thus the tables that we will consider are case-data tables.

1.2. CREATION OF CASES AS A COMPONENT OF DATA MODELING

The process of collecting, organizing, and analyzing data has been described by some as "data modeling" (e.g., Hancock, Kaput, & Goldsmith, 1992; Lehrer & Romberg, 1996; Lehrer & Schauble, 2000). We understand a model to be a simplification of a real-world phenomenon used for some purpose, such as for making predictions about, understanding, or controlling that phenomenon. Purpose is a necessary component of a model as it guides judgments about whether the simplifications have preserved the critical features of the real world. Accordingly, we share with these authors the view that modeling is a key aspect not only of producing statistical summaries, such as fitting curves to data, but of every component of the "investigative cycle" (Wild & Pfannkuch, 1999). Thus, even "deciding what data to collect, designing a structure for organizing the data, and establishing systematic ways of measuring and categorizing" involve modeling (Hancock et al., 1992, p. 339).

In this article we focus on the specific data-modeling step of recording observations, a step that results in documents such as Ives' case-data table. We want to consider recording observations separately from the work of creating measures, such as developing a meaning for, and method of determining, the maximum temperature of a day. Aspects of measure creation, and how they involve modeling, has been investigated by several researchers. Lehrer and Schauble (2000), for example, studied how young students developed ways of categorizing drawings made by children of various ages for the purpose of predicting the age of the child who made a drawing.

We propose that fundamental to making and recording real-world observations—the process of turning observations into data—is the conception of case. By case we mean the physical record of one repetition of a repeatable observational process. On January 1, 1985, Philip Ives went out to his weather station and wrote down values for the same set of observations he had made on roughly 13,000 previous days. Each repetition produced a case, which in Ives' weather data corresponded to a day. Cases occupy the rows of his table. When we can, we by convention begin a row with the value of an attribute that uniquely identifies the case. Accordingly, the left-most column of Ives' table contains the day of the month. If we were recording information about students in a class, we would most likely place student names at the beginning of the rows.

In creating cases, we map a real-world phenomenon into the data world. And because a case is an encoding of the observation of that object or event, and not an encoding of the object or event itself, cases are not just one but two levels removed from that real-world phenomenon. Being always aware that a case is a simplification, and a simplification of the observed event and not the event itself, is a big part of what guides expert exploration during data analysis.

David Moore's list of core components of statistical thinking begins with "The omnipresence of variation in process" (1990, p.135). Variation requires that there be a set of comparable entities (cases) that can be seen to vary. Thus, the transformation of observations into cases is at the very foundation of data modeling, because it is this process that creates comparable entities. When we make a graph displaying the distribution of our students' grades on the last test, an assumption underlying the sensibility of that display is that all of the values in it are in some sense equivalent and comparable. If we discovered our last grocery bill in the graph, we would remove it.

Ives' table makes it easy to see where one case ends and the next begins. It also conveniently places values that one might want to compare across cases, maximum temperatures, for example, in the same column. But this row-by-column structure is not only for convenience. It effectively binds information together, modeling the way in which the corresponding characteristics cohere in the real world where characteristics belonging to the same day or person are inseparable from that event or object. We see a man and can take notice of his hair color and height. As he moves around, these features move with him. If that last phrase seems odd it is because these features are integral to who he is. When we record values of our observations of this man's attributes into a table such as the one Ives made, we place values we obtained in a unique row, binding that information together so that it is clear to those who understand the conventions of this table that all these values belong to the same case. These values cohere to make a case; the observations have been transformed into data. Unlike the statement about the features moving with the man, it no longer seems odd to say that when we move a row in the table, all the values of the corresponding case move along with it. We point this out to highlight that a case is a simplification (or model) of its real-world referent. If in the process of transforming observations into data we fail to bind together information that belong to each entity we

are observing, we will most likely compromise our objectives. We call this property of a data model "case coherence."

In the new world of large data sets and analysis tools, the handling and structuring of data is becoming increasingly important. Given that data are now almost always analyzed with software packages, Wickham (2014) stressed the importance of having standardized ways of formatting data. The format that has become the virtual standard, and which he described as "tidy data," comprises a table where each column contains a variable (or attribute) and each row contains one observation (or what we call a "case"). In this article we refer to such tables as "flat tables." But in truly tidy data there is the further stipulation that a table contains only one type of observational unit. If you were collecting information about students in various schools, for example, you would make one table for the characteristics of the students (which would include the school they attended) and a separate table for characteristics of the schools, such as their size, location, etc. According to Wickham, this form provides "a standard way of mapping the meaning of a dataset to its structure" (p. 4).

1.3. PRIOR RESEARCH AND THE OBJECTIVES OF THE CURRENT STUDY

As we alluded to earlier, there is almost no research on how students go about recording and organizing data, either by hand or using software. In one of the only studies focused on students' facility with data recording and organization, Falbel and Hancock (1993) interviewed 13 students ranging in grades from 4 to 7 as they completed the "group separation problem" using the data-analysis software, *Tabletop*. These students had from 1 to 2 years experience working with versions of this software. Students were shown a target display in the software in which the names of four girls and three boys were separated into two different groups by gender. The students' task was to enter data into *Tabletop's* case table that would allow them to recreate the target display. *Tabletop's* case table was a conventional row-by-column structure as we have described. Thus the solution was to create a table with two columns, one labeled Name and the other Gender, and then enter each child's values into one of the seven rows. Most of the students created two columns, but these were labeled Boys and Girls. They then entered the students' names under the appropriate label resulting in four rows. This structure would not allow them to recreate the target display because *Tabletop* interpreted the rows as cases, which generally held the name of one boy and one girl. It took hints and considerable guidance before students could figure out and correct the problem, and even with this assistance, a few students were unable to find a solution.

One explanation for this result is that the students were not aware of the rigid conventions of *Tabletop's* case table, which demands that cases be entered as rows and attributes as columns. We have found similar problems even with experienced users of two similar tools we created, *TinkerPlots* (Konold & Miller, 2011) and *Fathom* (Finzer, 2012). In a typical scenario a teacher with a class of 12 boys and 15 girls will enter this information in the software's case table as shown in Figure 3 on the left. Expecting he has entered the information for 27 students, he is surprised when he makes the graph on the right to find apparently only one boy and one girl. Again, to enter this data in the way the software expects, the teacher would need to create 27 rows and just one column named Gender.

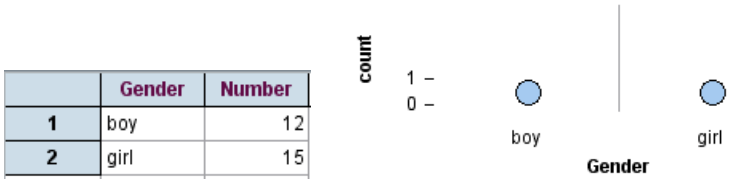


Figure 3. A user enters the frequency of boys and girls in the table on the left then is surprised when only two dots appear in the graph of the data on the right; this table expects its rows to be cases, so it "thinks" the first row is an individual "boy" who for some reason has the Number "12"

Falbel and Hancock (1993) described the table most students made as a "set-based" structure, because it divided individuals into different sets according to the value of one of their properties. The structure that *Tabletop* wanted they described as a "property-based" structure because it consisted of cases, each with the values of their properties. As an alternative to the "rigid convention" explanation for their findings, the researchers posited that a property-based structure may be more complex conceptually than the comparable set-based one. This is because the property-based structure requires that students conceptualize properties or attributes (e.g., Name and Gender) to which they can then assign values. So they cannot just think "John, male." They need to think "Name, John; Gender, male." We should point out, however, that the set-based table was perfectly adequate in terms of representing or modeling the external world by binding information to cases. Few people would have any difficulty recovering from the set-based structure these students made the fact that it contains information about seven students, one of whom was Sally, a female.

What motivated us to further pursue this research on students' facility with recording and structuring of data was a project we were working on together that involved creating a new data-analysis environment, CODAP, to handle data generated during the play of computer games. We were developing these materials as part of the Data Games project, a collaborative effort funded by the National Science Foundation. The games produced data of two types: data associated with particular moves or actions that a student made while playing one game, and data about each game, such as the game number and score. These data could have been recorded in a "flat" row-by-column table like the one Ives used for recording the daily weather values. But this would have had various limitations. Accordingly, we decided to record and display the data in a hierarchical structure (see Figure 9 for an example). Given the known difficulties just referred to with students and teachers understanding and using flat case tables, we were concerned that students might have difficulty understanding and working with what we assumed would be a more complex data structure.

Note that Ives did not use a strictly row-by-column table to record his observations. He recorded the month and year of his observations outside of the table in the upper corners of each page, and this is critical information. He could have added two columns to his table, one for year and one for month, and entered values for each case there. But this would have increased his recording time and required extra space. In writing them once at the top, he trusted that his readers would understand that all of the cases in the table inherit these values.

Applying our definition of case, we claim that Ives created cases of two different types. One case type was a day, and those were recorded as rows in the table. The other case type was a month, and those corresponded to separate pages in his notebook, one page for each

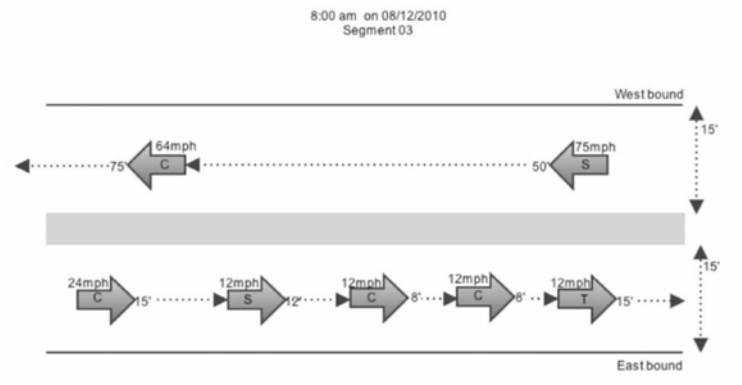
month. (In fact, you might claim that he created cases of three types, the third being a half a month.) Attributes associated with months included the month's name, its year, and averages for the month, which Ives included at the bottom of each page (not visible above).

Our primary objective in this research was to see how students would record and organize the data in a rather complex situation. For that purpose, we created snapshots depicting traffic moving along sections of a highway and asked students in both individual interviews and group settings to develop a way of recording all of the information available on the snapshots. We hoped in this way to learn 1) whether students could successfully bind information to cases, and if so, 2) what methods they used to accomplish this, and 3) whether these methods changed with the age of the students.

2. METHODS

2.1. TASK

We introduced the task by telling participants that city planners were "studying the traffic along roads that lead into and out of the city" and that as a part of that study they were "collecting data at various times of the day along short road segments." We then presented them with the snapshots shown in Figure 4 and told them that the snapshots "include all of the information the planners want collected." The vehicles are depicted in the snapshots as arrows with the type of vehicle (**T**ruck, **C**ar, **S**uv) inside the arrow, speed above the arrow, and following distance in front of the arrow. We pointed to and identified each piece of critical information that they were to record, which included the vehicle information just mentioned along with time, date, segment number, lane width, and direction, for a total of eight attributes. We then gave participants a blank sheet of paper on which they were to record the data and told them their "data sheet should not be a drawing of these snapshots. It should be an organized record of the data values on the snapshots." We also told them that we might later give them more snapshots and have them add data from those to their data sheet. The mention of this possibility was intended to encourage the development of an organization that could readily accommodate additional data.



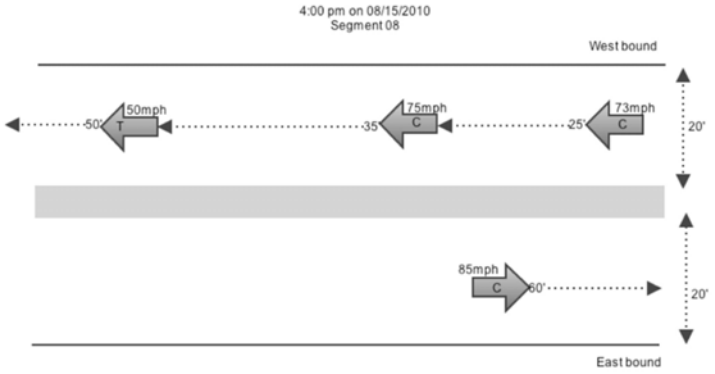


Figure 4. Snapshots showing traffic along two road segments

We intentionally did not specify any questions the city planners had because our pilot studies suggested that this could influence the way participants organized the data or lead them to record only those data needed to answer the particular questions we gave as examples. This could easily be cleared up in an interview setting but not in the group-administered setting. We also were careful not to suggest that the data be recorded in a specific format, a table for example. We added the instruction about not making a drawing of the snapshots after observing many students doing this in our early pilot testing. We elected to use schematics for the snapshots rather than actual photographs, for example, because we wanted to focus on the task of recording and organizing data rather than that of developing coding and measurement schemes, which a photograph would have required.

2.2. PROCEDURE

We administered this task in two formats: group administration and individual interviews. In the group-administration format, we gave the instructions verbally to the class and displayed the snapshots on an overhead projector. After giving the instructions orally, we passed out packets that included the task instructions we had read, the two snapshots, and a blank piece of paper. After ten minutes, we collected their work.

In the individual interviews, we instructed participants to "think aloud" as they completed the task, and we gave them a warm up problem so they could practice this technique. After this, the instructions paralleled those in the group administration format. As they worked on the task, interviewer probes were limited to prompts to think aloud and requests to describe what they were doing.

After interview participants had finished constructing their data sheets, we removed the snapshots and asked them what questions they imagined the city planners might have. Then we presented them, one at a time, six short questions (see Appendix A) about the traffic (e.g., "What percentage of the vehicles are cars?"). For each question, we asked participants to show us how they would get the answer to the question from their data sheet. The primary purpose of these questions was to make participants aware of any deficiencies in their recording schemes without explicitly pointing them out. After we had given them all six questions, we showed them the snapshots again and gave them the opportunity to redesign their data sheet or add additional information to it.

Finally, we showed interview participants two data sheets (see Appendix B) supposedly made by other students and asked them whether they could use those data sheets to answer the same six questions and whether they saw any advantages or disadvantages of these data

sheets compared to the organization they had used. One sheet was a "flat" case-by-attribute table, with vehicles as the case. The other was a three-level hierarchical table with two large cells for date and time, and contained within those the two lanes with direction and width, and nested within those the vehicle information. We presented first the data sheet that the interviewer judged to be the most different from the one the participant made.

There was no fixed time limit for the interviews, though they generally ran about 30 minutes. The interviews were videotaped and transcribed.

2.3. PARTICIPANTS

We conducted individual interviews with 22 students (9 middle school and 13 high school). The 162 participants in the group administration consisted of 69 middle school students, 21 high school students, 52 college students, and 20 teachers. The college students were undergraduates at a large California university enrolled in a course on the theory and practice of sampling. The middle and high school students were in six different classes in which we were working as part of the Data Games project. We administered the task at the beginning of our work with them, before they played the games or used the data-analysis environment. The 20 teachers were participants in workshops associated with the same project.

3. RESULTS AND DISCUSSION

In this article, we focus primarily on analysis of the data sheets participants made. We will use excerpts from the interviews to elaborate on our analysis of the data sheets but will not present here a comprehensive analysis of those interviews. For the interviewed participants, we consider the data sheets they initially created, not their revisions in response to follow-up tasks and probes. We reasoned that these initial data sheets would be most comparable to the ones made by participants in the group-administered task. Additionally, in almost all of the interviews, participants on their first pass created structures that successfully bound cases. Thus revisions they did make involved recording information that they missed on the first pass, such as lane width, and adding this additional information did not require them to develop an alternative structure.

We report many of the results broken down by three age groups: Middle School (MS), High School (HS) and Adults (a combination of the college students and teachers.)

3.1. CODING THE DATA SHEETS

We developed coding schemes to capture various aspects of the data sheets, including the type of representation used to encode the data, how the information was organized, and how complete it was. These codes immerged over time as we examined and compared the various data sheets. Two coders independently coded all of the data sheets and negotiated disagreements to end up with a single set of codes for all data sheets.

To determine inter-rater reliability, we randomly selected 59 data sheets (originally 60, but one turned out to be a second page of another) and set them aside before we began developing our coding schemes. Once we had settled on a final set of coding categories and had refined them, we independently coded these 59 data sheets, which we had not previously seen. The measures of reliability, which we will report below, were computed using Cohen's Kappa, which gives the proportion of agreement between the two raters after accounting for chance agreements.

3.2. COMPLETENESS

Some data sheets did not have enough on them to warrant inclusion in our analysis. To weed them out we established criteria for "completeness." The traffic snapshots included information on eight attributes. Only 33% of the data sheets (17% MS, 41% HS, 46% Adults) included every bit of the available information on the snapshots. But many who participated in the group administration format appeared to have run out of time, as we only allowed ten minutes. Also, omission of some attributes (such as lane width), seemed to be oversights rather than an inability to figure out where to record that information on the data sheet. Thus we established a less rigid criterion for completeness to allow inclusion of data sheets that had information on the most important attributes and that were organized clearly enough that we could reasonably infer the structure participants were using and thus how they would have encoded any missing information. This required that information on the vehicles in at least one lane was recorded and included direction, speed, vehicle type, and following distance.

By this criterion, 83% of the data sheets were complete (72% MS, 88% HS, 93% Adults). We eliminated from further analysis the 31 data sheets (22 from the MS group) that we deemed incomplete. Inter-rater reliability for coding 5 levels of Completeness was 0.748, as determined by Cohen's Kappa.

3.3. BINDING

One of the reasons we incorporated as many attributes as we did in the traffic snapshots was to increase the challenge of developing a structure that could hold all the data values and preserve their relationships. As we argued in the introduction, to do this requires establishing a coherent case identity and a structure that binds case information together. We did not code for binding separately. It was one of many categories we developed to characterize the "Structure" of the data sheets. We describe Structure below, along with information about inter-rater reliability.

Given the relative complexity of the data, we expected that a large percentage of at least the middle school students would not be able to successfully encode the data by binding case information together. Rather, we expected many would encode the information in independent pieces, losing the information about the relationships among pieces. Figure 5, a portion of the data sheet of a middle school student, shows an example of this. This student listed the vehicle types and speeds together, but then coded the following distances in a separate line, not indicating which distances belonged to which vehicles. (The value of 25' for distance actually belongs to C = 73 mph.) Contrary to our expectations, data sheets using this type of unbound organization were rather rare, occurring in only 8 cases, all middle school students (or 14% of them).

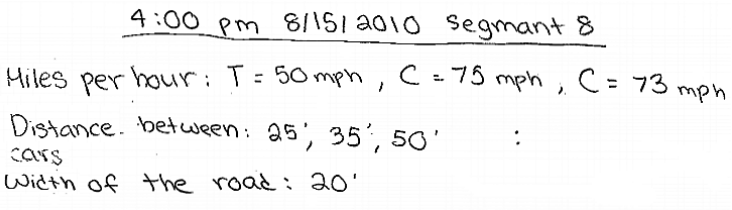


Figure 5. A portion of the data sheet of a middle school student in which the values for "Distance between cars" are not bound to the vehicle cases in the line above

3.4. SUCCESS

Considering as successful completion of the task only those data sheets that were both complete and bound, the overall success rate was 79%. By this criterion, even 62% of the middle school students were able to create a bound structure that could hold the critical information from the snapshots. This percentage was probably affected by a comment made by the teacher in a 7th grade class who whispered that she was not sure how students would do since they had not covered *graphing* yet. Students could hear the aside, and we believe as a result 7 of the 29 students attempted to encode the data by making a graph, an approach that otherwise was rather rare. Six of these students were unable, and understandably so, to encode in a graph enough of the information to be considered complete. The success rate for the high school students was 88% and for the adults 93%.

For all the analyses that follow, we have eliminated incomplete or unbound data sheets. We have also removed two middle school data sheets that were reproductions of the snapshots. We were left with 145 data sheets: 48 MS, 30 HS, and 67 Adults.

3.5. TYPE OF REPRESENTATION

We categorized the 145 "successful" data sheets as employing either a table, graph, or narrative to hold and organize the traffic data. The measure of inter-rater reliability (Cohen's Kappa) for coding Type of Representation was .921.

In the task instructions, we were careful not to specify the type of representation; in particular, we did not use the word "table." While 1 participant encoded the data successfully in a graph, the majority (74%) recorded the information in the form of a table or multiple tables as shown in Figure 6.

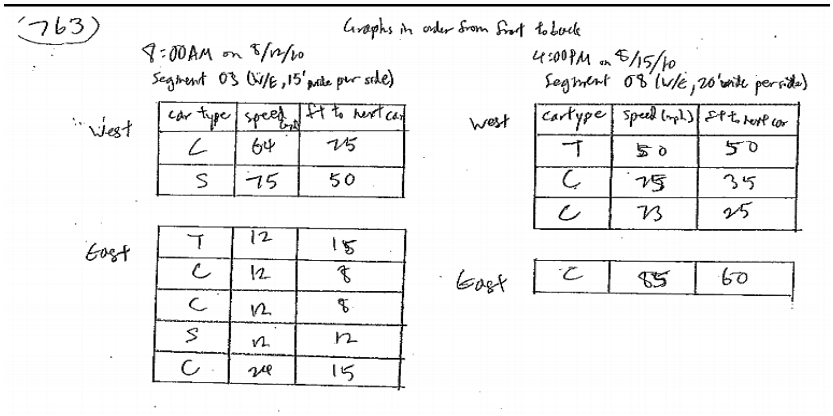


Figure 6. Data sheet of a high school student who used several tables to encode the data; time, date, segment number, lane width, and direction function as table headings

Twenty-five percent of the data sheets used a narrative style to encode at least part of the data. Figure 7 shows an example narrative from a middle school student. We coded a representation as a narrative when at least some of the information was encoded as sentences or phrases. Here, the student encoded the vehicle types, speeds, and following distances using clipped sentences: "Car 4 going 73 mph, 25 ft behind Car 5..."

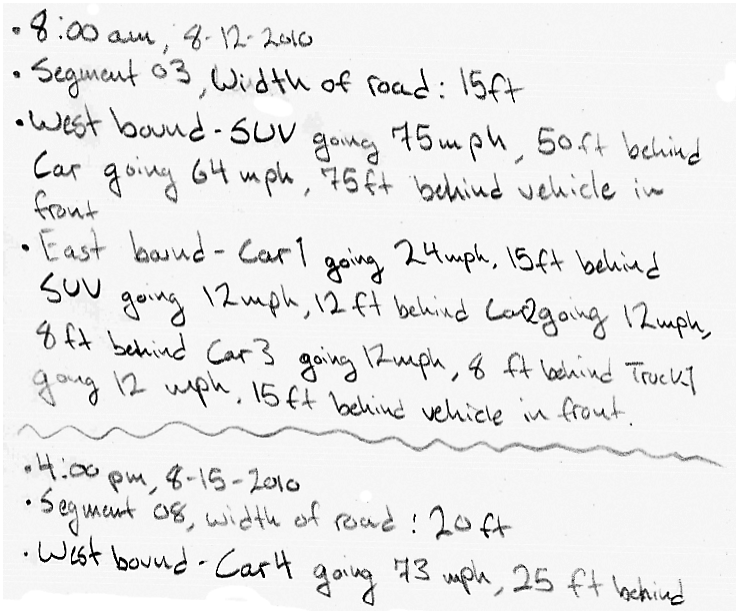


Figure 7. Portion of a data sheet made by a middle school student using a narrative style

As Table 1 shows, the percentage of particular types of representations varied according to age group. As the age of the participants increased, narratives decreased and tables became the preferred form.

Table 1. Percentage of age group using each type of representation

	Type of Representation		
Age Group	Graph	Narrative	Table
Middle school	1%	60%	38%
High school	0%	17%	83%
Adult	0%	3%	97%

3.6. STRUCTURE

We developed a method of capturing both the specific information contained in data sheets as well as how it was organized, which we call "Structure." This involved listing the attributes encoded on the data sheet, describing their spatial arrangement and their function. By function we mean, for example, using an attribute (or value of an attribute) as a heading, which you can see in Figure 6 and which we describe in more detail below.

The method we developed to code Structure was equally applicable to tables, narratives, and graphs. Cohen's Kappa for coding Structure was .807. We will not report results at this level of detail, as this coding resulted in a total of 120 unique descriptors for the data sheets we coded. Rather, we used these Structure codes as the basis for forming higher-level categories, such as Binding as described above.

3.7. TYPES OF TABLES

The tables that participants created were of two general types: flat and hierarchical. But many of these also made use of headings—values of attributes such as the directions "East" and "West"—to separate the data into sections. We describe these various categories of tables below.

Flat tables without headings Twenty-six percent of the tables overall were what we term "flat tables," which made no use of headings. Figure 8 shows a portion of the flat table produced by a middle school student. Each row is a separate vehicle, with their attributes listed along the top. Note, too, that the first listed attribute is the type of vehicle, the attribute most closely associated with the individual case identities.

Car Type	Speed mph	Direction	Distence from car in frount	Time	Segment	DatE	road width
SUV	75mph	W	50ft	8:00am	3	8/12/2010	15'
Car	64mph	W	75'	8:00am	3	8/12/2010	15'
Car	24	E	15'	8:00am	3	8/12/2010	15'
SUV	12	E	12'	8:00am	3	8/12/2010	15'
car	12	E	8'	8:00am	3	8/12/2010	15'

Figure 8. A portion of the flat table produced by a middle school student. Many students, like this one, used the term "car" not just as one of the vehicle types but also as the super-category instead of "vehicle."

As we mentioned earlier, most statistical software, including Fathom and TinkerPlots, require that data be entered in this form. Given how ubiquitous this format is in the data world, we expected that a majority of our Adult group would used this form. But only 32% of the tables of the Adult group were flat tables. For the middle- and high-school students, the percentages were 22% and 12%, respectively.

Flat tables with headings The most common approach was to organize the information as multiple flat tables grouped under headings. Overall, 58% of the tables created were of this variety; by age group the percentage were 67% MS, 84% HS, and 46% Adults. The data sheet in Figure 6 provides a prototypical example. First, the data were organized into two sections based on the time, date, and segment number. Within these two divisions, data were further clustered according to the two lane directions, east and west. This resulted in 4 flat tables, which contained the type, speed and following distance of the 11 vehicles.

We refer to elements like "east" and west" as well as the times, dates and segment numbers written along the top as "headings." Headings are values of attributes. "East" and "West" are values of the unnamed attribute "Direction;" "8 am" is a value of the unnamed attribute "Time." In addition to organizing the data into subgroups, headings also serve to encode all of the information listed within the cluster they create. Thus, we know that all of the vehicles listed in the table at the upper left of Figure 6 were heading west at 8 am, etc. Headings are an efficient way to encode information because all of the cases under a heading inherit its value. Rather than use year and month as attributes in his table (see Figure 1), Philip Ives wrote them as headings and spared himself considerable work. Similar to the student who created the data sheet in Figure 6, Ives did not identify the attribute by name in his headings by writing, for example, "Year: 1985," or "Month: January." From the context, it is easy to infer what "January" and "1985" refer to.

Hierarchical tables without headings Nine percent of the tables (all made by Adults) we consider "hierarchical tables." In hierarchical tables, all of the attributes are structured as columns and all the cases as rows. But the cases are at more than one level, with the relationships among cases at different levels indicated by a nested spatial arrangement.

In the example in Figure 9, there are four cases at the higher level that we might consider to be the four road lanes, their time and date of observation, the segment they are part of, and their direction and width. Vehicles, with their type, speed, and following distance, are nested within the appropriate lane row.

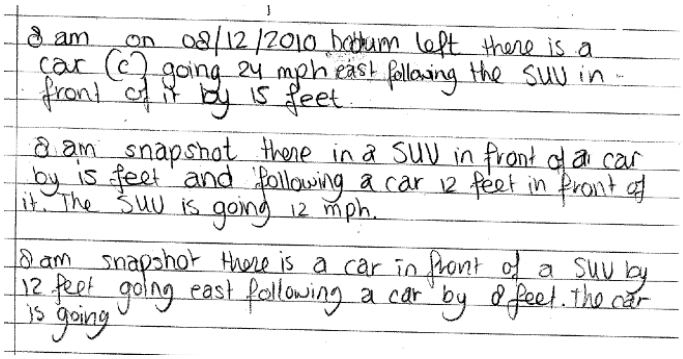
Time & Date	Segment Number	Direction of traffic	Width of Lanes	Type of Vehicle	Speed (mph)	Distance to the vehicle in front (in ft)
8:00am 8/12/2010	03	westbound	15'	SUV	75	50'
				car	64	75'
8:00 am 8/12/2010	03	Eastbound	15'	car	24	15'
				SUV	12	12'
				car	12	8'
				car	12	8'
				truck	12	15'
4:00 pm 8/15/2010	08	Westbound	20'	car	73	25'
				car	75	35'
				truck	50	50'
4:00 pm 8/15/2010	08	Eastbound	20'	car	85	60'

Figure 9. Hierarchical table without headings made by a college student. Narrow rows on the right are vehicle cases, which are nested within four thicker rows, corresponding to the four lanes

Hierarchical tables with headings Five percent of the tables were hierarchical tables with headings. These headings served the same function as we described above for flat tables. They created subsets of the data, where each subset had its own hierarchical table and the values in the headings applied to all cases in that subset. There were 5 such data sheets, 3 made by Adults, 1 by a high school student and 1 by a middle school student.

3.8. NARRATIVE STRUCTURES

All but one of the 36 narratives also used headings to encode part of the information. The narrative in Figure 7, for example, used clipped sentences to encode the vehicle types, speeds, and following distances. But this vehicle information was separated into groups with headings that specified the time, date, segment number, road width, and direction. The only "flat narrative," which made no use of headings, was produced by a college student who repeated all values for each vehicle (see Figure 10).


8 am on 08/12/2010 bottom left there is a car (c) going 24 mph east following the SUV in front of it by 15 feet.

8 am snapshot there in a SUV in front of a car by 15 feet and following a car 12 feet in front of it. The SUV is going 12 mph.

8 am snapshot there is a car in front of a SUV by 12 feet going east following a car by 8 feet. the car is going

Figure 10. A portion of a flat narrative by a college student in which all the information, including time and date, is repeated for every case (vehicle)

We argued that a row-by-column data table such as Ives created, binds information to cases. This binding relies on a convention that specifies that all information in a row belongs to a case. And even if a person did not explicitly know this convention, because the rows resemble one another, with similar values occurring in the same column, most people looking at the table would correctly interpret each row as a day. We might say that the cases are well-formulated and thus easily recognized. In narratives, cases are bound by grammatical conventions and in particular the understanding that information in a sentence generally refers to the subject of the sentence. We speculate that many of the students who used narratives rather than tables did not immediately see how they could bind all the critical information using tables, but certainly knew how to use sentence and paragraph structures (including headings) to accomplish that.

In this sense, they were able to encode cases, as we have defined them. However, many of the narratives that students created would not have served the intended purpose of the planning board who wanted to make use of this information. Consider, for example, the narrative in Figure 11. Here, the items appearing as separate bullets are not cases but rather pieces of information. What binds the information is both the order and connecting words such as “Also.” We can recover information from this data sheet about each vehicle, but it takes considerable work. There are no indicators of where one case ends and another begins, no sense of a repeated observation.

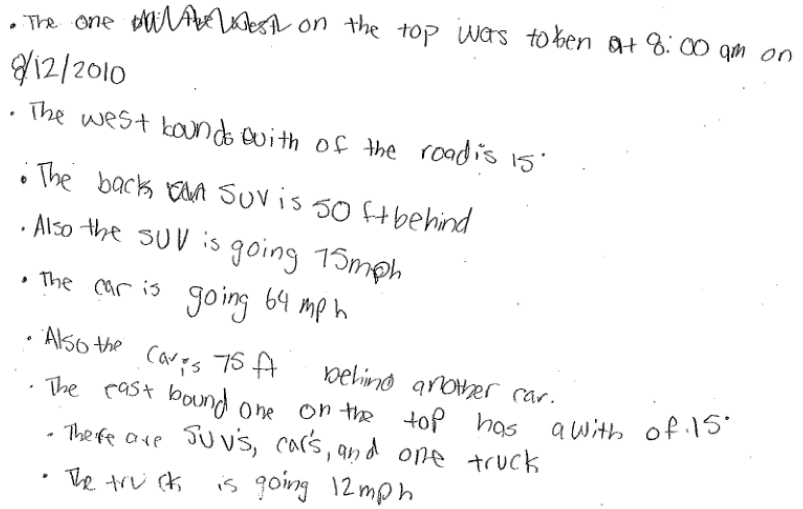


Figure 11. A portion of narrative-style data sheet made by a middle school student in which the cases are difficult to see

Contrast this with the narrative in Figure 12. Here, each clipped sentence is headed by the value that identifies it as a vehicle. There is also an attempt to align each sentence to create implicit columns of comparable information. The unit of repetition (the case) is clearly discernable. Nearly the only narrative part of this data sheet is the word "behind." In the first line, the student uses a rather long phrase to make it known that the identity of the vehicle ahead of the truck is unknown. It may have been these complexities, which the student believed needed to be preserved, that led to the use of a narrative rather than a table. Thus one hypothesis about why many students used narratives rather than tables in this study is that they could not see how to maintain some of the necessary information in other than narrative form. In terms of modeling, they believed critical information would be lost if they used a table.

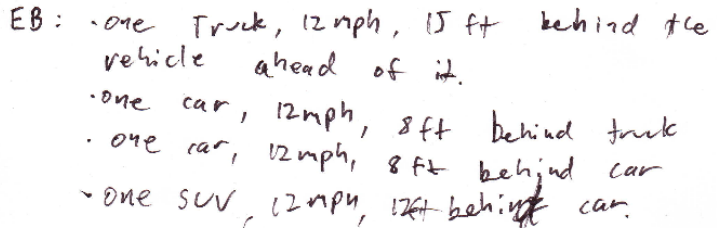


Figure 12. Portion of a narrative-style data sheet from a middle school student in which the cases are clearly distinguishable

There were two instances of students who began encoding the information as a narrative then switched to using a table, as seen in Figure 13.

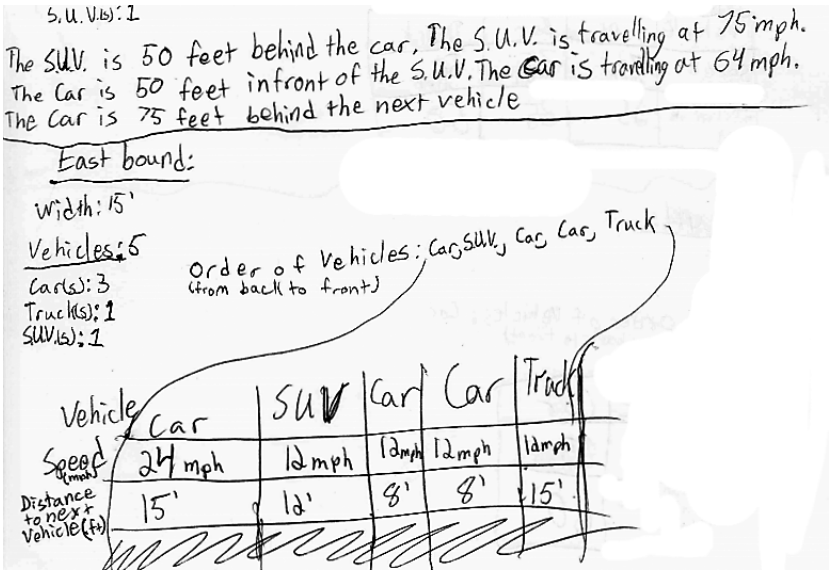


Figure 13. Portion of a data sheet from a middle school student who transitioned from a narrative to table format; although by convention cases are placed in rows, this student placed cases in columns and the attributes in rows

Narratives ordinarily do not require that cases be conceived in terms of attributes and their values, or what Falbel and Hancock (1993) referred to as "property-based." A vehicle can be thought of simply as "a car going 24 mph" rather than as "a vehicle of type car and speed 24." But note when that same information is encoded in a table, as seen in Figure 13, attribute names appear. Without them, it would be difficult to know what some of the values were (the following distances in particular). This raises a second hypothesis—that younger students in this study tended to use narratives because narratives do not require thinking in terms of attributes and values, whereas the corresponding tables do. This lends support to the explanation offered by Falbel and Hancock for the difficulty they observed in their students—that it is conceptually more complex to think of data as "property based," i.e., as attributes along with their values. Lehrer and Schauble (2000) made similar observations in analyzing the different approaches taken by 1st and 2nd graders vs. 4th and 5th graders as the students worked to develop methods for distinguishing drawings made by other students of various ages. With a little support, most all of the older students adopted criteria that made use of dimensional, attribute-value criteria in making their judgments, while none of the younger students adopted such methods, relying instead on "holistic perceptual judgments" communicated in detailed narratives.

3.9. USING NESTED STRUCTURES INCLUDING HEADINGS TO DISTINGUISH AMONG DIFFERENT TYPES OF OBJECTS

We had initially expected that the flat table would be the most common type of structure used to encode the data. We had not anticipated the frequent use of headings. We believe that, for the most part, participants were using these headings not just for efficiency in encoding but also to model the situation accurately. The snapshots contain vehicles on road segments, and at different observation times. Many of the students' data sheets used one set of headings for date and time, under those a set for lane directions and widths, and under those a list (or narrative) of vehicle data. With cases at three different levels, they modeled

the structure of the real-world situation. To create a single flat table involves creating cases of one type—vehicles. In this flat table, all attributes must be assigned to vehicles, including those such as "direction" that might more naturally be thought of as belonging to the lane. This flattening preserves the necessary information but seems to require a different way of modeling and thinking about the objects. Lanes, in a sense, disappear as objects in their own right, and their attributes become attributes of vehicles. In the interviews we find some evidence that this sort of transformation in thinking occurs. As some participants moved between considering structures of different types, the way they talked about various attributes changed.

Below is a portion of the interview transcript of a high school student. He had created a narrative using headings, which suggested three types of objects: 1) road segments along with their time and date of observation, 2) lanes, which had a width and direction, and 3) vehicles, with their type, speed and following distance. Below is how he described these objects as he was creating his data sheet:

> Well, on these 15' lanes, there's one [lane] with two and one with 5 cars, and on the 20' wide lanes there's one with three and one with one car. So it seems that the thinner the lane is, the more cars are going to be on it. ... I would note which lanes are going westbound and which are going eastbound.

After he had created his data sheet and answered various questions about it, the interviewer showed him the flat table shown in Appendix B and asked him to interpret it.

> There's a column for the date. It seems like by each individual vehicle, so there's like the date and the time it was photographed or whatever. There's the road number it was on, like segment 3 and 8, the direction it was going, so east or west, the width of the lane, the type of vehicle, the speed, and the distance from the car in front of it.

In interpreting the flat table, he no longer talked about the direction of the lanes. Instead he talked about the direction *vehicles* were going. Indeed, he now described the attributes, including time and date of observation, and the road segments, as attributes that belong to "each individual vehicle."

Our study supports the view that in encoding data, students attempt to create a structure that models the target phenomenon as closely as they can. In this example, they in general attempt to construct cases that correspond to the different sorts of objects they perceive in the real world. To do so with these data, they most often use headings. Less frequently do they flatten the world into a single set of objects and even less often do they construct a hierarchical table without headings.

4. CONCLUSION

As the data revolution gains momentum and the data we have to analyze become greater in size and complexity, the importance of understanding how to structure those data for analysis becomes increasingly important. In this regard, the results of this study are encouraging. They suggest that the majority of students, even those with little or no experience structuring data, have some basically sound ideas about data and data structure. Indeed, given the complexity of the data in the traffic snapshots and the fact that learning to record and organize such information is not currently a part of the mathematics or science curricula, we were surprised by the high percentage of representations, even among middle school students, that could hold all the data and preserve the critical relations among them. Most of our participants seemed to appreciate that the kinds of questions that would be important to address in this context would require that the numerous values involved be bound to cases. And while many of the middle school students did not seem to know how

to construct tables that would do this job, most of them understood the value of organizing their narratives around cases and to make this case-structure apparent to the reader.

A major motivation for undertaking this research was our concern that students might have even more difficulty working with the hierarchal data structures we were building into our new software, CODAP (Finzer & Damelin, 2016), than we had observed them having with the flat structures of virtually all data analysis software, including *Fathom* and *TinkerPlots*. To our surprise, students were more likely to create nested data structures than they were to produce one flat table, which suggests to us that hierarchical structures, which involve a type of nesting, might in fact be more intuitive and easier to interpret than flat tables.

Based on this study we explored the possibility of allowing users to make use of headings in structuring data in a case-table with the hope that by allowing headings, students would encounter fewer difficulties than they currently do when they need to input data into a software tool using the strict attribute-value convention. Unfortunately (and in accord with Wickham, 2014), we had to abandon this plan as we discovered that it was not possible to write a set of rules (algorithm) for deciding whether a word or phrase at the top of a column, say, was an attribute name or a heading. Without such rules, students would need to be able to specify whether the label over a column was a "heading" or an attribute, and this would pose more difficulties than just requiring that all columns be labeled with attribute names.

In the current study, we allowed students to use whatever methods they desired to record and organize the data. But such flexibility is not possible in a computer environment, at least not currently. As a follow-up to this study, we are investigating how students understand data that have been structured for them in hierarchical form in CODAP. One might think that this would be an easier task than creating such a structure on one's own. However, the challenge with using a structure that has already been built is coming to understand the conventions used to create it, and those conventions are likely to be different than the ones the students would have employed on their own. Our preliminary finding, as well as our experience working with students in classrooms, is that students have less difficulty working with hierarchically structured data than they do with the corresponding flat structures. We believe that at least part of the reason for this is that hierarchical structures more closely resemble and model the real-world objects that students are trying to explore.

ACKNOWLEDGEMENTS

Rick Gaston, Vishakha Parvate, and Jamie Stevenson, helped develop the task and conduct some of the interviews. Deb Nolan consulted with us on the task and arranged for the collection of data from the university students. This research was supported by the National Science Foundation under Grant Nos. DRL-0918653, DRL-0918735 and DRL-1316728. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Ehrenberg, A.S.C. (1977). Rudiments of numeracy. *Journal of the Royal Statistical Society*. Series A (General), *140*(3), 277-297.

Estrella, S. (2014). *The table object: An epistemological, cognitive and didactic study*. English translation of *El objeto tabla: un estudio epistemológico, cognitivo y didáctico*.

Unpublished doctoral dissertation. Pontificia Universidad Católica de Valparaíso: Chile.

Falbel, A., & Hancock, C. (1993). Coordinating sets properties when representing data: The group separation problem. In I. Hirabayashi, N. Nobda, K. Shigematsu, & F. Lin (Eds.), *Proceedings of the 17th Annual Meeting of the International Group for the Psychology of Mathematics Education,* (vol. *2*, pp. 17-24). Tsukub, Japan: University of Tsukub.

Finzer, W. (2012). *Fathom dynamic data software* [Computer Software]. Emeryville, CA: Key Curriculum.

Finzer, W., & Damelin, D. (2016, April). *Design perspective on the Common Online Data Analysis Platform (CODAP).* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Friel, S. N., & Bright, G. W. (1998). Teach-Stat: A model for professional development in data analysis and statistics for teachers K-6. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching, and assessment in grades K-12* (pp. 89-117). Mahwah, NJ: Lawrence Erlbaum.

Hancock, C., Kaput, J., & Goldsmith, L. T. (1992). Authentic inquiry with data: Critical barriers to classroom implementation. *Educational Psychologist*, *27*, 337-64.

Konold, C., & Miller, C., D. (2011). *TinkerPlots: Dynamic data exploration.* Emeryville, CA: Key Curriculum. Available from http://www.tinkerplots.com

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

Lehrer, R., & Schauble, L. (2000). Inventing data structures for representational purposes: Elementary grade students' classification models. *Mathematical Thinking and Learning*, *2*(1&2), 51-74.

Martí, E., Garcia-Mila, M., Gabucio, F., & Konstantinidou, K. (2011). The construction of a double entry table: A study of primary and secondary school students' difficulties. *European Journal of Psychology of Education*, *26*(2), 215–235.

Moore, D. S. (1990). Uncertainty. In L. A. Steen, (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, D.C.: National Academy Press.

Prodromou, T. (2015). Students' emerging reasoning about data tables of large-scale data. *International Journal of Statistics and Probability*, *4*(3), 181.

Robson, E. (2003). Tables and tabular formatting in Sumer, Babylonia, and Assyria, 2500 BCE-50 CE. In M. Campbell-Kelly, M. Croarken, R. Flood, & E. Robson (Eds.), *The history of mathematical tables: From Sumer to spreadsheets* (pp. 1-47). Oxford University Press: Oxford.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1-23.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-265.

CLIFFORD KONOLD
Scientific Reasoning Research Institute
Middlesex House Rm. 129
111 Country Circle
University of Massachusetts Amherst
Amherst, MA 01003-2545
USA

APPENDIX A. INTERVIEW QUESTIONS ABOUT THE TRAFFIC DATA

We posed these six questions to interview participants after they had finished constructing their data sheet. Their task was to determine whether they could answer each question from their data sheet and to indicate the information they would use from that sheet to answer the question.

1. What percentage of the vehicles are cars?

2. What is the average speed of vehicles in the westbound direction at 4 pm?

3. Is there more traffic overall heading eastbound or westbound?

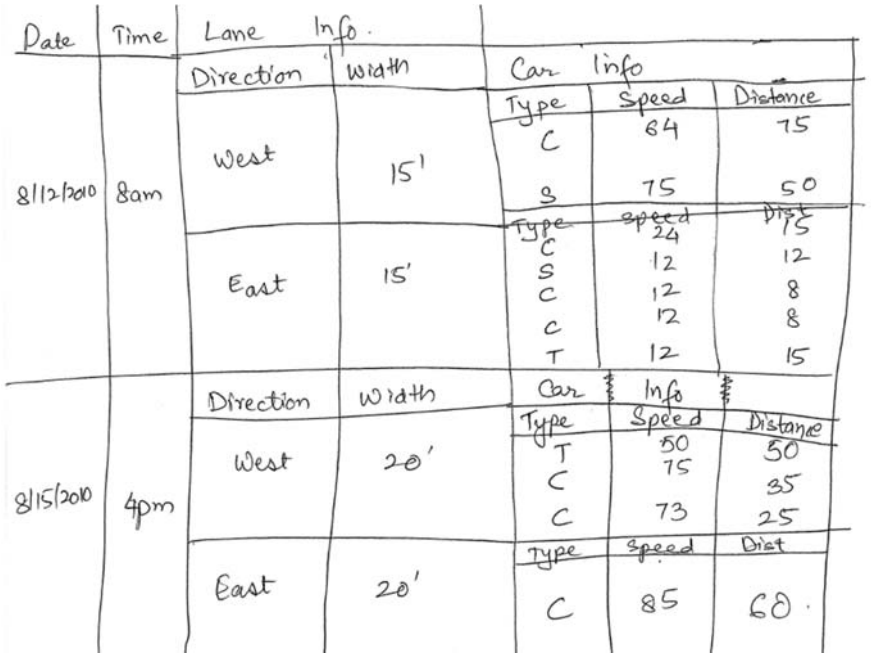
4. Are trucks in general slower than other vehicles?

5. Does how fast vehicles are going have anything to do with the distance between vehicles?

6. Does the number of vehicles on the road have anything to do with the width of the road?

APPENDIX B. SAMPLE DATA SHEETS SHOWN TO INTERVIEW PARTICIPANTS

Example of a data sheet using a hierarchical table.



Date	Time	Lane Info.		Car Info		
		Direction	Width	Type	Speed	Distance
8/12/2010	8am	West	15'	C	64	75
				S	75	50
		East	15'	~~Type~~ C	~~speed~~ 24	~~Dist~~ 75
				S	12	12
				C	12	8
				C	12	8
				T	12	15
		Direction	Width	Car Info		
				Type	Speed	Distance
8/15/2010	4pm	West	20'	T	50	50
				C	75	35
				C	73	25
		East	20'	~~Type~~ C	~~Speed~~ 85	~~Dist~~ 60

Example of a data sheet using a "flat" table.

Date	Time	Road#	Direction	Width	Veh Type	Speed	Distance
8/12/2010	8am	3	West	15	C	64	75
8/12/2010	8am	3	West	15	S	75	50
8/12/2010	8am	3	East	15	C	24	15
8/12/2010	8am	3	East	15	S	12	12
8/12/2010	8am	3	East	15	C	12	8
8/12/2010	8am	3	East	15	C	12	8
8/12/2010	8am	3	East	15	T	12	15
8/15/2010	4pm	8	West	20	T	50	50
8/15/2010	4pm	8	West	20	C	75	35
8/15/2010	4pm	8	West	20	C	73	25
8/15/2010	4pm	8	East	20	C	85	60