Research on Student Understanding of Data Organization

How do they work with complex and unstructured data?

Fall 2009-Fall 2012

Research Objectives

As part of the Data Games project, we are researching how students record and organize multivariate data. This research is informing the design of new software interfaces for Fathom and TinkerPlots that will allow students to explore and understand data that live in other than “flat” data structures — the structures that most software tools currently limit themselves to.

We have designed the Traffic Problem to explore the following questions:

1. What methods do novices and experts use to systematically record data with multiple attributes?
2. In recording data, do students employ a recognizable notion of “case”?

Task and Materials

City planners are studying the traffic along roads that lead into and out of the city. As part of the study, they are collecting data at various times of the day along short road segments. Here are two such road snapshots.

The snapshots include all of the information that the planners want collected: the time and date ... So on the 8:00 a.m. snapshot at the top left is a car (C) going 64 mph following 75 feet behind the vehicle in front. Following behind it is an SUV. ...

Research Protocol Summary

On a blank sheet of paper, record the data from these two snapshots. Your data sheet should not be a drawing of these snapshots. It should be an organized record of the data values on the snapshots. ... Later, we may give your completed data sheet to other students who will use it to answer several questions about the flow of traffic along this road. These students will not see the snapshots you saw. Therefore, your data sheet needs to include all of the information these student will need to answer the questions we give them.

Revision Cycles

In the process of refining the Traffic Problem, we have given various versions of it to 31 novices and 6 experts. Most of the novices were high school students.

Preliminary Analyses

Our initial assumption was that novices would not record and organize data in a way that would allow them later to answer questions about relations among attributes. Surprisingly, however, the majority of them created forms that housed all the information in the snap shots and organized the data in a way that maintained all the important relations. Specifically:

- About half organized data in “flat” tables.
- A majority used an organizational method that kept the information about individual vehicles together in such a way that it is possible to determine, e.g., the correlation between distance and speed.
- Many employed an organizational method consistent with a hierarchical data model, in that they partitioned information spatially to reflect different case levels (date/time, lane direction, vehicle information).

We are developing coding schemes that attempt to capture the brilliance of novices and their desire to use it to answer several questions about the flow of traffic along roads that lead into and out of the city. As part of the study, they are collecting data at various times of the day along short road segments. Here are two such road snapshots.

Student Work from Early Versions of the Protocol

In the process of refining the Traffic Problem, we have given various versions of it to 31 novices and 6 experts. Most of the novices were high school students.

Software Implications

A prototype of nested data tables. The nesting of attributes within levels addresses the desire of many novices to record data efficiently so as to avoid the seeming needlessness of repetition of, e.g., date and time for every vehicle. But it also allows analysis at different levels of case without actively restructuring the unit of analysis.

Acknowledgments

The Data Games project wishes to thank Doug Page and his students at Galileo Academy of Science and Technology; Josh Zucker, Lick-Wilmerding High school; Lisa Huang and project staff at KCP Technologies and UMass Amherst, for their valuable support of this research work.