Data Transformations: Restructuring Data for Inquiry in a Simulation and Data Analysis Environment

Michelle Hoda Wilkerson, UC Berkeley; mwilkers@berkeley.edu
Kathryn Lanouette, UC Berkeley; kathryn.lanouette@berkeley.edu
Rebecca L. Shareff, UC Berkeley; becca@berkeley.edu
Tim Erickson, eeps media, eepsmedia@gmail.com
Nicole Bulalacao, UC Berkeley; nbulalacao@berkeley.edu
Joan I. Heller, Heller Research Associates, jheller@gordonheller.com
Natalya St. Clair, Concord Consortium, talya@concord.org
William Finzer, Concord Consortium, wfinzer@concord.org
Frieda Reichsman, Concord Consortium, freichsman@concord.org

Abstract: We explore *data transformations*, actions investigators take to make datasets more useful for intended inquiries. Fourteen young adults were interviewed while they interacted with online science and engineering games and were provided their own gameplay log data to improve their scores. We investigate the conditions under which data transformations are likely to emerge; provide examples of data moves as enacted by participants; and propose an initial taxonomy of data transformations and potential developmental supports.

Working with data is an important epistemic practice across disciplines, including within K–12 education. However, students often treat data as a static representation of "the answer", rather than as a source of evidence. Specifically, we are interested in how learners come to engage with provided or "second-hand" data (Hug & McNeill, 2008) as transformable (Duschl, 2008). Using provided data has the potential to engage learners with topics and scopes of analysis that would be difficult in a classroom setting. But students may not automatically see those data as relevant for inquiry, or engage in critical reasoning about those data. Thus, we argue that just as students should learn to construct, represent, and analyze their own data in service of an investigation (Konold & Pollatsek, 2002), they should also learn how to re-construct, re-represent, and re-analyze provided data.

Theoretical framework

To engage in data transformation, learners must navigate three nested, interactive processes (Figure 1). The first involves identifying a goal and deciding whether provided data are *relevant* for that goal. The second involves considering the *context* in which provided data were generated, including the data collection instruments used, for what purposes the data were constructed, and how well the data reflect their own experiences of the context. The third concerns making data more useful through restructuring, supplementing, or subsetting, constructing new measures or visualizations, and so on. Such *reconstructions* are made possible, and reified as valuable, by an emerging genre of data analysis software (e.g., Tableau, RStudio) used by practitioners and increasingly the general public (Wickham, 2014). The Common Online Data Analysis Platform (CODAP) used in this study makes reconstructions accessible to K-12 students. These nested processes are similar to others' descriptions of evidence construction and use in the science education literature (e.g., Kelly, Regev, & Prothero, 2007).

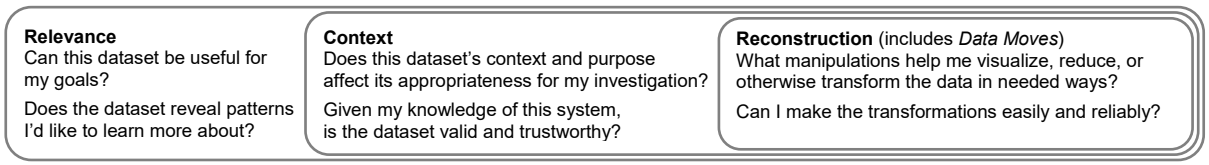

Figure 1. Data transformation includes determining the relevance, context, and required manipulations of data.

Methods

Data Science Games (IIS-1530578) seeks to develop students' data transformation skills through games about core science and engineering ideas. Each game, and the data it generates, is embedded within CODAP, which allows players to visualize, organize, and manipulate their data. The data require some degree of transformation before they are useful for making decisions or understanding the game's underlying principles. In this study, Stebbins (Figure 2, left) is designed to engage learners with basic ideas underlying natural selection through analyzing the convergence of data toward a favorable trait. BARTy (Figure 2, right) is designed to engage learners in optimizing engineering solutions for a large transit system (Bay Area Rapid Transit, or BART), by exploring

patterns in nested data structures.

We conducted guided cognitive interviews with 14 public high school and community college students as they played Data Science Games. We captured synchronized video of students' computer screen activity and audio-video records of their comments and interactions with one another and the interviewers. Several team members engaged in co-viewing and interpretive coding of each interview video (Jordan & Henderson, 1995). After several rounds of analysis, we focused specifically on moments in the interviews at which the need for data moves arose (for example, a student suggests data should be represented or structured differently), or when data moves were actually executed by participants within the CODAP environment.

Figure 2. Stebbins (left) focuses on natural selection; BARTy (right) focuses on public transit ridership data.

Results

Across all interviews, participants only engaged in data reconstruction once they identified the relevance of the data for addressing some need, and had made enough sense of the data context. This led to vastly different treatments of data across interviews: *data rejected as irrelevant*, *data for inquiry about gameplay*, and *data for inquiry about phenomenon*. Participants also reconstructed data in different ways. Re-representing data was common; fewer engaged in re-structuring or re-scoping data. For example, Figure 3 demonstrates the back-and-forth nature of data transformation with a focus on re-representation. Here, a participant (Pam) chose to explore the relationship between color and score (3, 4). A mismatch between the data and her experience (6) led her to recognize the dataset as inappropriate (7); she then answered her question with a more appropriate dataset (8).

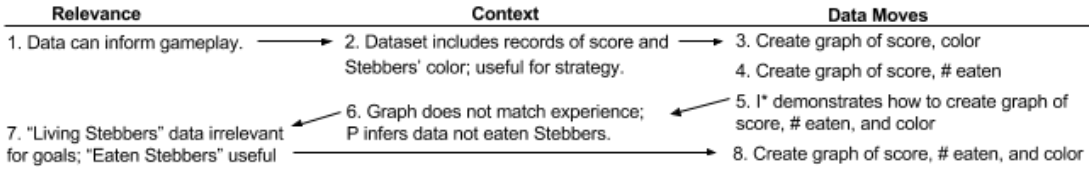

Figure 3. Summary analysis of Pam's engagement with data; classified as *data for inquiry about gameplay*.

The analyses demonstrate how particular data moves are deeply situated within specific goals, and are shaped by learners' consideration of data context. It is important for learners to recognize data as a resource for inquiry, and data transformation is an important part of this process. This project represents an initial step toward understanding the contexts in which students learn to manipulate large datasets.

References

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education*, 32(1), 268-291.

Hug, B., & McNeill, K. L. (2008). Use of first-hand and second-hand data in science: Does data type influence classroom conversations? *International Journal of Science Education, 30*(13), 1725-1751.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *Journal of the Learning Sciences, 4*(1), 39-103.

Kelly, G. J., Regev, J., & Prothero, W. (2007). Analysis of lines of reasoning in written argumentation. In S. Erduran & M. P. Jimenez-Aleixandrae (Eds.), *Argumentation in Science Education* (pp. 137–158). Dordrecht, The Netherlands: Springer.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259-289.

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

Acknowledgments

This material is based on work supported by the National Science Foundation under Grant IIS-1530578.