DIVING DEEPER INTO CATEGORICAL DATA

Jan Mokros, Jacob Sagrans, Andee Rubin, Traci Higgins, and Ada Ren
Science Education Solutions and TERC
JMokros@scieds.com

The purpose of this research is to examine how 11–14-year-old students examine and make sense of authentic datasets that include two or more categorical variables. We created an afterschool program in which students explored existing data using the Common Online Data Analysis Platform (CODAP). Individual think-aloud interviews were conducted, during which students were questioned about relationships between variables while they used CODAP. We describe how students engaged in exploratory data analysis that involved comparing two or more categorical variables of interest, including complex comparisons in a large contingency table. This study is one of the first to examine understanding of categorical data that are not pre-structured.

INTRODUCTION

Our work focuses on how students interrogate and make inferences about public datasets generated by large government surveys. In the research reported here, we examine how middle school students examine relationships between categorical variables, including variables that have three or more levels. This type of data analysis is usually not done until students are at a later stage in their schooling, but we posit that it is both possible and beneficial for younger students to be engaged in the examination of patterns in categorical data. Because we work with datasets that are relevant to topics that are interesting to students aged 11–14 (middle school in the United States), the statistical issues we focus on emerge from the interaction between students and the data and are somewhat unpredictable. We started with the datasets themselves, rather than statistical concepts that we wished to teach. This meant we were forced to deal with the complexities of the categories the original investigators chose to use. The datasets we have incorporated in one of our curriculum modules (“Injuries on and off the Field”) includes mainly categorical variables with multiple levels, which is common in many large datasets.

As we describe in our literature review below, most of the research on reasoning about relationships among categorical variables deals with variables that have two levels (often yes/no or male/female), or possibly three. By contrast, some of the categorical variables in the datasets we use have 10–15 possible values, which makes different and more complex kinds of data-based reasoning possible. This led us to examine how students work with this complexity and what strategies they would employ to make sense of the data.

Most research on students’ understanding of relationships between categorical variables presents students with contingency tables, not with the individual cases used to construct the table. Because the students we worked with were using the Common Online Data Analysis Program (CODAP; Concord Consortium, n.d.), they created representations of individual cases in a grid, isomorphic to a contingency table but with a different visual impact and with the opportunity for students to choose to look at cases, frequency counts, or percentages. Through interactions with these representations, we were able to learn how students made sense of the case frequencies and what lenses they brought to their exploration of the data. Much of their reasoning, we believe, followed from their engagement in the construction of the representations. Importantly, the students were also aware of other variables in the dataset and often included references to these in interpreting the graph.

LITERATURE REVIEW

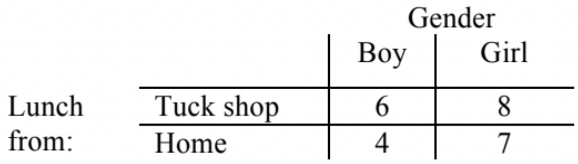
According to the newly revised Guidelines for Assessment and Instruction in Statistics Education II (GAISE II) framework, middle school students should learn to work with categorical data and should become “comfortable describing the manner in which [categorical] data are organized in two-way [contingency] tables as well as noticing the benefits a visual representation can provide” (Bargagliotti et al., 2020, p. 52). More generally, students in grades K–12 should have opportunities to explore patterns of association between two categorical variables.

Although there has been considerable research on how students use data to make comparisons (for a review, see Biehler et al., 2018), much of this work has been in the context of examining datasets that include both numerical and categorical variables or only numerical variables. There also has been

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today’s Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

significant research on how students understand contingency tables, beginning with Inhelder and Piaget’s (1955) research with adolescents, showing that understanding associations in a table requires understanding proportionality, probability, and combinatorics. More recent research continues to shed light on how people make sense of two-way contingency tables and also examines how people think about covariation more generally; this process is referred to by Garfield and Ben-Zvi (2008) as “covariational reasoning.” Garfield and Ben-Zvi identify several significant challenges people face in reasoning about covariation, including students’ tendencies to let prior beliefs influence their reasoning; focus on certain cells in contingency tables more than others; have greater difficulty in understanding inverse correlations (as opposed to positive ones); and draw causal inferences where none may exist.

Research in the last decade has examined the challenges described above in a more nuanced way. Natural language can be ambiguous about which variable is being “percentaged” in a contingency table (Knapp, 2015). It can be challenging to match a particular question with the way one thinks about conditional probability. For example, Budgett and Puloka (2019) posed the following question to both students and experts: “Who is more likely to get lunch from a tuck shop, boys or girls?” and showed them the contingency table in Figure 1 below.



		Gender	
		Boy	Girl
Lunch from:	Tuck shop	6	8
	Home	4	7

Figure 1. Contingency table from Budgett and Puloka (2019, p. 6)

Most experts concluded that boys were more likely to get lunch from a tuck shop because 60% of boys but only 53% of girls got lunch from a tuck shop. However, one statistician concluded that girls were more likely to get lunch from a tuck shop because 8 of the 14 tuck shop customers (57%) were girls. This statistician’s answer corresponds to a slightly different question, namely, “Is a customer of a tuck shop more likely to be a boy or a girl?” Clearly, language matters.

Context and beliefs also matter when interpreting contingency tables. Beginning in the 1990s, a number of studies have employed a contingency table where one variable is smoking (yes/no) and the other is lung disease (yes/no). Typically, the number of smokers in the table is larger than the number of nonsmokers, but the proportion of each group who has lung disease is identical, so there is no covariation between the two attributes. Watson and Callingham (2015) developed a rubric to show levels of students’ thinking about these data and concluded that students’ knowledge of the dangers of smoking interfered with making a correct conclusion about the lack of association in the table.

A considerable amount of research has been devoted to understanding how students utilize additive vs. multiplicative reasoning (or proportional reasoning more generally) in the context of contingency tables. In a study with young students (7–10 years old), most employed additive reasoning when using contingency tables to compare the number of chips of different colors in unequal-sized bags (Obersteiner et al., 2016). About 14% of the older students were able to use multiplicative or proportional reasoning and explain how they did so. Students had a hard time verbalizing their proportional reasoning, though it was easier for them to do so when there were simple ratios in the table. Saffran et al. (2019) found that students who examined a series of two-by-two contingency tables were better able to use proportional reasoning to explain correct conclusions that had been provided to them, compared to using proportionality to justify their own conclusions. The authors found that explanations involving ratios were rare and that part of the reason may be due to limited working memory.

Technology can support students’ understanding of covariation. Much of the research in support of this claim focuses on how students investigate the relationship between numeric variables and employs technology that preceded the development of CODAP; the examples included in Biehler et al.’s (2018) overview are from Cobb’s mini-tools (Cobb et al., 2003) and TinkerPlots (Konold & Miller, 2005). Part of our interest in focusing on relationships between categorical variables is curiosity about how CODAP might provide additional possibilities for students as they explore categorical variables (see the example in Figure 2 below).

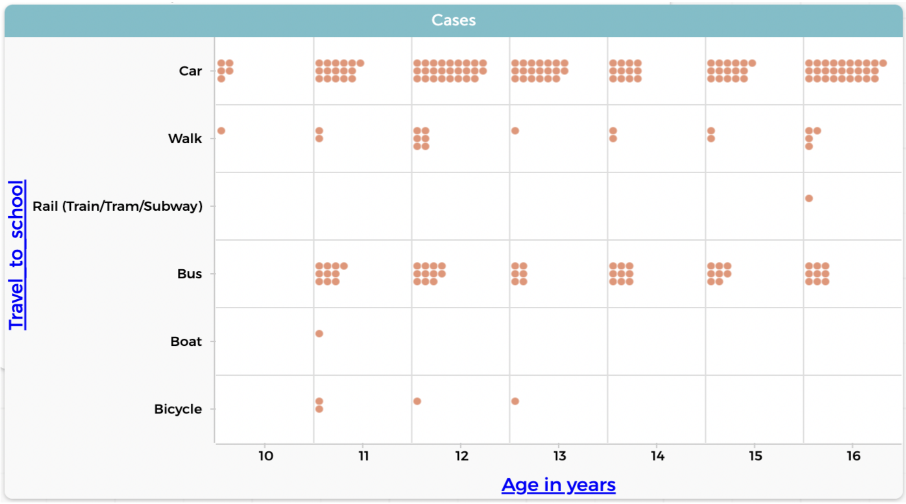


Figure 2. CODAP graph showing how students travel to school versus age in years in a Census at School (https://ww2.amstat.org/censusatschool/) dataset, with age treated as a categorical variable

In sum, the literature shows that it is quite challenging for people of all ages to reason about relationships between categorical variables. Language often gets in the way, as when people start a sentence with "Certain people are more likely to do X…" but fail to finish the sentence with "compared to Y" or "compared to a different group of people." Sometimes the challenges are linguistic and at other times they are perceptual and relate to the visualizations that subjects are given or make on their own. Data visualizations highlighting one variable versus the other can switch one's conclusions about relationships in ways that are analogous to the figure/ground shifts in perception relating to visual illusions like the one involving either a vase or two faces. Radford (2014) discusses how visual perception may interfere with mathematical generalization, postulating that the learner's eye must become "domesticated" to looking in the right places. In the case of categorical data, students must make decisions about how to look at the data—by cells, by rows, or by columns—as they make different kinds of comparisons.

THE CONTEXT: DATA CLUBS

The Data Clubs (n.d.) project, a National Science Foundation-funded Collaborative Research Project, is based in both urban Massachusetts settings and in rural areas of Maine. The goal is to introduce data science to middle school students in out-of-school settings (afterschool programs and summer camps), with a focus on students historically under-represented in STEM (i.e., students of color, girls, and rural students). We partnered with community organizations to help us reach these students, including nonprofit organizations, school districts' afterschool programs, and summer camps. Due to schedule constraints of these settings, we designed 10-hour modules that can be offered flexibly, from one hour per week over 10 weeks to a more intensive offering of 10 hours over the course of one week or two. Students use CODAP, designed by Concord Consortium (n.d.), as their data visualization and analysis tool.

Each of three modules focuses on a topic (as opposed to particular statistical techniques) and includes multiple publicly available datasets that our team has curated and modified to make suitable for our audience. All of the students who we report upon below had participated in the "Injuries On and Off the Field" module, using data from the National Health and Nutrition Examination Survey (NHANES) (Centers for Disease Control and Prevention, 2016) and the National Health Interview Survey (NHIS) (Centers for Disease Control and Prevention, 2018).

METHOD

The major research question was: "How do students use CODAP to visualize and reason about the relationship between two categorical variables, with at least one variable having many (e.g., 10) levels?"

After completion of the module, each student was asked to participate in a 35–45-minute one-on-one virtual interview. Here, we consider data from 13 students who were interviewed during the later phase of the project after we had identified a core set of robust interview questions about relationships between two or more categorical variables. These students volunteered to participate in an afterschool program in a town in coastal Maine. Every student in this program was interviewed. Within the interview, students engaged in several data investigations. Each interview began by orienting students' attention to the structure of the dataset they would be exploring (rows, columns, cases, and the definition of variables). If students could not remember how to make graphs or make other moves in CODAP, we reminded them how. First, students examined a simpler dataset from Census at School containing 200 cases relating to "travel to school." The variables included three categorical ones, including mode of transportation for getting to school and age, as shown in Figure 2. We asked students to explore a question of their own and also asked them a standard question about whether mode of travel to school differed for younger and older students.

Then, students investigated variables found in the NHIS dataset, which was familiar to them because it was used in the Injuries module. They were first asked to explore a relationship between two variables that they found interesting. If they had not chosen two categorical variables, they were then asked a standard question: "How would you investigate the relationship between the type of injury and whether the injury resulted in a trip to the emergency room, using the tools in CODAP?" Students made their own graphs to address both the standard question as well as additional comparative questions that they found interesting. Video recordings of the interviews were transcribed, including all verbal interactions as well as the work students did in CODAP to make graphs, compute summary information (such as counts or percentages), manipulate the dataset, and point out comparisons using their cursors.

ANALYSIS AND PRELIMINARY RESULTS

We focus here on this question: How did students use CODAP to explore the injuries dataset as they investigated the relationship between type of injury (with 10 levels) and whether an injury resulted in admittance to the emergency room (two levels)? All students we interviewed made graphs similar to the one in Figure 3 by dragging one of these attributes onto the horizontal axis and the other onto the vertical axis.

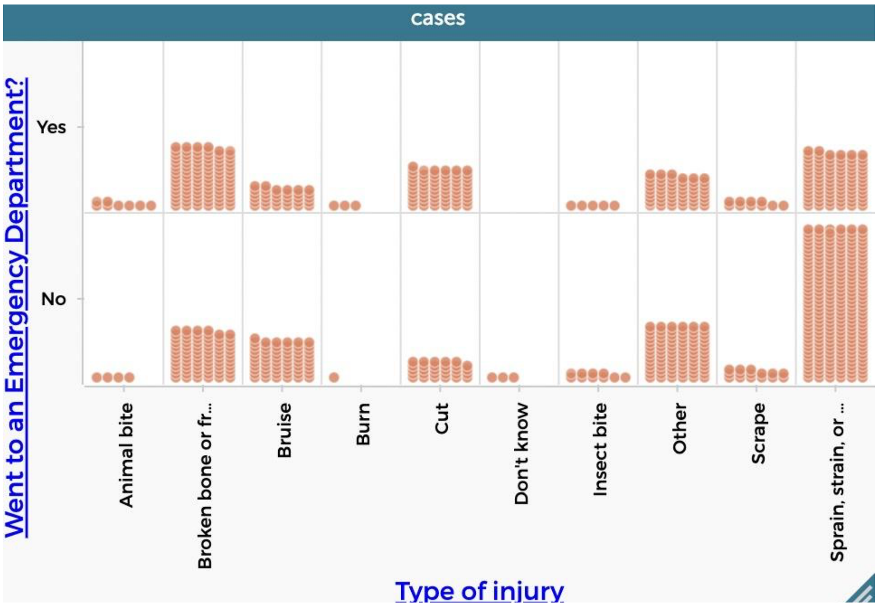


Figure 3. CODAP graph showing went to emergency room (yes or no) versus type of injury

We found that there were five general ways in which students examined the relationships between these categorical variables. However, all of the students interviewed used at least two of these strategies to examine the relationships that could be extracted from the above graph. The five major strategies students employed were as follows.

1. Focus on a single case. Example: Joey looks at the cell for people who were cut but did not go to the emergency room and mouses over the cases (dots) within the cell. This move reveals the ages of individual cases (via pop-up boxes that appear in CODAP). He says: "Well the 85-year-old [who was cut] didn't even go to the emergency room."
2. Focus on the cell with the largest frequency. Example: Liz looks at the entire graph and says, "I'm seeing a lot of people did not go to the emergency room for a 'sprain, strain, or twist.' That is, by far, like my biggest number, so my eyes are drawn to it right away."
3. Examine just one row or column. Example: Cassidy looks at the small category of "burn" and compares going versus not going to the emergency room, saying: "25% didn't [go to the ER] and 75% did. Like I feel like a burn is either like a not that bad injury or like a pretty bad injury and it doesn't happen that often."
4. Examine within row or within column patterns, looking at several examples. Example: Theo examines frequencies for each of several injuries, using CODAP's count tool, saying, "I see broken bone or a fracture is at 94 [cases that went to emergency room] where sprain, strain, or twist is 86 [cases that went to the emergency room]."
5. Coordinate rows and columns while looking for patterns. Example: Zach says, "I think that most like sprain and insect bites and cuts, most of these categories, they all kind of have like the similar proportions of yes and no [for went to the emergency room], like kind of around like 40 to 60% [uses CODAP's percentage tool with columns]. There are a few outliers, I think, especially with like sprain and maybe cuts as well."

Students spent significant time engaged in data investigation, sometimes exploring more than one question and adopting more than one approach. For example, a student often would start by noticing a cell with a large number of injuries (such as fractures). Then, they might use CODAP's percentage tool to compare the cells representing the proportion of fractures that resulted in a trip to the emergency room versus those that did not. Many students noted that a surprisingly high number/percent of fractures did not involve going to the emergency room. At that point, they might move on to another injury, complete the same process, and then compare that injury with their initial analysis of fractures. They did not always move from simple reasoning to more complex reasoning. Intriguing individual cases, or cases that resembled their own injury experiences often drew them in at various points in their analysis, but they were readily able to switch their attention back to the bigger picture, especially when reminded of the major question they were investigating.

CONCLUSIONS

Although analysis of interviews is still underway, we are finding that the types of reasoning students are using seem different from those reflected in the literature on contingency tables. Although the literature seems consistent on the difficulties that students and adults have with two-by-two tables, we are seeing that students exhibit successful approaches in dealing with the relationship between a categorical variable with many levels and one with two levels in the context of using CODAP as a visualization tool. When one of the variables has a large number of values, such as "type of injury" we are seeing a kind of reasoning similar to ways in which students examine a numerical distribution, i.e., seeing the "shape" of the distribution, noting the modal clumps, and pointing out holes. When doing exploratory analyses to address questions of these data, students freely used CODAP tools in no particular order to help them see patterns. For example, students often rearranged the categories, ordering them by the relative frequency of different types of injuries. They went back and forth between using percentages and frequencies, depending on what they wanted to find out. As instructors, we did not teach students a particular sequence of moves for examining the data, so it is not surprising that students came up with their own unique pathways for data exploration.

What might these insights mean for the order in which we introduce students to thinking about categorical variables? Most textbooks spend a long time on two-by-two contingency tables, usually with the longer-term goal of preparing students for using the chi-square statistic. Only later do they move to variables with more possible values, assuming that it is simpler to analyze a two-by-two table than one with more rows and/or columns. But is it really simpler to start with two-by-two tables? We have found that students themselves almost universally chose to investigate complex relationships about two-by-many categorical variables. In fact, a number of them examined three-way relationships between categories, which CODAP facilitates through color coding. Much of students' work in these interviews consists of exploratory data analysis that would be recognizable to a statistician, and we believe that this exploration is an important foundation for later understanding of categorical data.

REFERENCES

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K–12 guidelines for assessment and instruction in statistics education II (GAISE II). A framework for statistics and data science education.* American Statistical Association; National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Biehler, R., Frischemeier, D., Reading, C., & Shaughnessy, J. M. (2018). Reasoning about data. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds), *International handbook of research in statistics education* (pp. 139–192). Springer. https://doi.org/10.1007/978-3-319-66195-7_5

Budgett, S., & Puloka, M. (2019). Making sense of categorical data—Question confusion. In S. Budgett (Ed.), *Decision making based on data. Proceedings of the Satellite conference of the International Association for Statistical Education.* ISI/IASE https://iase-web.org/documents/papers/sat2019/IASE2019%20Satellite%20114_BUDGETT.pdf?1569666564

Centers for Disease Control and Prevention. (2016). *National health and nutrition examination survey (NHANES) 2015–2016.* https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015

Centers for Disease Control and Prevention. (2018). *National health interview survey (NHIS) (1997–2018)*. https://www.cdc.gov/nchs/nhis/1997-2018.htm

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13. https://doi.org/10.3102/0013189X032001009

Concord Consortium. (n.d.). *Common Online Data Analysis Platform (CODAP)* (Version 2.0) [Computer software]. http://codap.concord.org

Data Clubs. (n.d.). *About data clubs.* TERC. https://www.terc.edu/dataclubs/about-data-clubs/

Garfield, J. B., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). *Developing students' statistical reasoning: Connecting research and practice*. Springer. https://doi.org/10.1007/978-1-4020-8383-9

Inhelder, B., & Piaget, J. (1955). *De la logique de l'enfant á la logique de l'adolescent* [From the logic of the child to the logic of the adolescent]. Presses Universitaires de France.

Knapp, T. R. (2015). "Percentaging" contingency tables: It really does matter how you do it. *Research in Nursing & Health, 38*(4), 323–325. https://doi.org/10.1002/nur.21666

Konold, C., & Miller, C. (2005). *TinkerPlots: Dynamic data explosion* (Version 3.0) [Computer software]. Learn Troop Pty Ltd. https://www.tinkerplots.com/

Obersteiner, A., Reiss, K., & Bernhard, M. (2016). How do primary school children solve contingency table problems that require multiplicative reasoning? In C. Csikos, A. Rausch, & J. Szitanyi (Eds.), *Proceedings of the 40th Conference of the International Group for Psychology of Mathematics Education* (Vol. 3, pp. 387–394). PME.

Radford, L. (2014). The eye as a theoretician: Seeing structures in generalizing activities. *For the Learning of Mathematics, 30*(2), 2–7.

Saffran, A., Barchfeld, P., Alibali, M. W., Reiss, K., & Sodian, B. (2019). Children's interpretations of covariation data: Explanations reveal understanding of relevant comparisons. *Learning and Instruction, 59*, 13–20. https://doi.org/10.1016/j.learninstruc.2018.09.003

Watson, J., & Callingham, R. (2015). Lung disease, indigestion, and two-way tables. *Investigations in Mathematics Learning, 8*(2), 1–16. https://doi.org/10.1080/24727466.2015.11790348