STUDENTS' REASONING ABOUT ASSOCIATION OF CATEGORICAL VARIABLES

Stephanie Casey[1], Rick Hudson[2], and Lauren Ridley[3]
[1,3]Eastern Michigan University, Ypsilanti, MI, USA 48197
[2]University of Southern Indiana, Evansville, IN, USA 47712
scasey1@emich.edu

Statistical association is an important concept in statistics. An exploratory study examined how students reason about statistical association of categorical variables using both numerical and graphical representations. Task-based interviews were conducted with thirteen students ages 11 to 13 prior to formal instruction. When prompted to make a graph to display the conclusions they reached numerically, successful students made side-by-side pie charts or segmented bar graphs, but most students were unable to produce a meaningful graph to compare the variables. When later asked to interpret previously constructed graphs, they were most successful with segmented bar graphs and least successful with eikosograms. When students failed to interpret the graph correctly, they often interpreted the percentages with incorrect referents. These results have curricular and software implications.

INTRODUCTION

Statistical association is a fundamental statistical idea in school curricula (Burrill & Biehler, 2011). According to curriculum standards from around the world (e.g., Australia: Australian Curriculum, Assessment and Reporting Authority, 2013; Brazil: Ministério da Educação, 2017; U.S.A.: National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), students ages 11-13 learn to assess association of two categorical variables. Previous studies completed by mathematics educators (e.g., Batanero, Estepa, Godino, & Green, 1996) and psychologists (see Garfield & Ben-Zvi, 2008, for a review of such research) have concluded that humans' assessment of data for association is largely inaccurate; thus, it is an area where proper education is vital.

Effective teaching supports students in building advanced knowledge from their prior understandings (Bransford, Brown, & Cocking, 1999). For this to occur, educators (including teachers, curriculum authors, and software developers) need to know students' common conceptions regarding a topic. In fact, this knowledge is considered an essential component of mathematical knowledge for teaching (Ball, Thames, & Phelps, 2008). Thus, this important exploratory study examined how students ages 11-13 (ages at which they are first learning this material in school) reason about association of categorical variables. This paper presents our preliminary results from the study focused on students' creation and interpretation of graphs.

METHODOLOGY

Thirteen students (8 males, 5 females) ages 11-13 voluntarily participated in the study at the end of the 2017 school year. They had not received formal instruction about how to analyze whether two categorical variables are associated. Each participated in a 30- to 50-minute, semi-structured, two-part interview conducted by an author. The first part of the interview engaged the students in analyzing the Vehicles dataset with CODAP software (see https://codap.concord.org/releases/latest/static/dg/en/cert/index.html#shared=29571), allowing the students to choose two categorical characteristics of vehicles they thought might have a relationship then assisting them in creating a two-way table with frequencies and percentages displayed to analyze whether this was the case. The last task in this portion of the interview asked students to sketch and label a graph that could be used to convince someone of the relationship (or lack thereof) they just discovered then describe what they would say about the graph.

The second portion of the interview began with the presentation of a new dataset regarding the characteristics of 33 different granola bars. Next, a series of five pre-made graphs regarding these granola bars were shown to the students (see Figure 1) which included three different graph types: side-by-side bar graph, segmented bar graph, and eikosogram (also known as a ribbon chart, it is a variation on a segmented bar graph where the width of the bars corresponds to the relative frequencies in the categories. Three of the graphs showed variables that were associated and the

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

remaining two showed no association. Each time a graph was displayed, the corresponding two-way table was also shown.

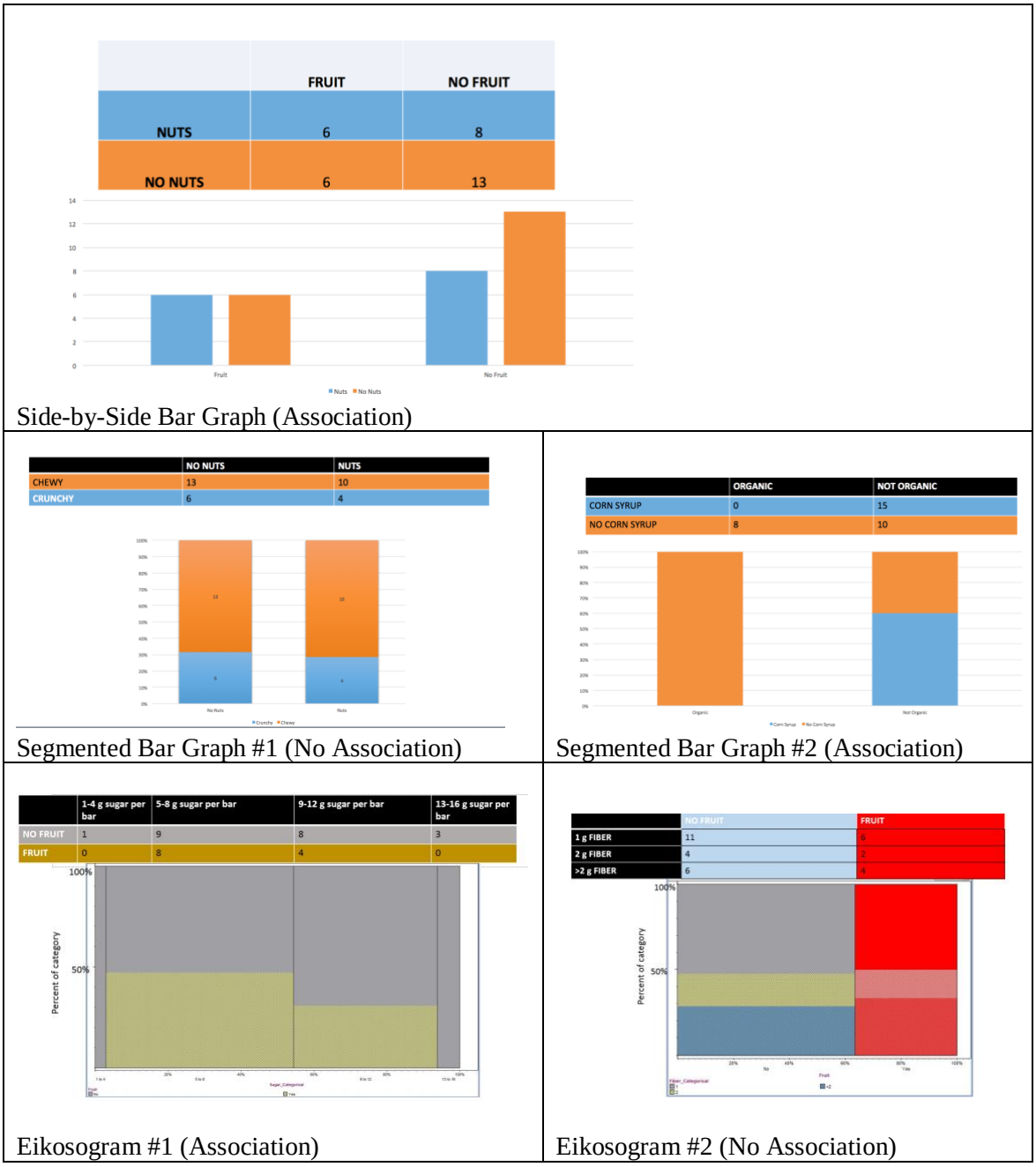


Figure 1. Graphs for Granola Bar Data Set

The students were introduced to the first graph of each type in random order then they analyzed the second segmented bar graph and second eikosogram also in a random order. They were asked to interpret each graph to determine if there was a relationship between the variables displayed (the variables were stated by name for each graph).

The interviews were documented with screencasts, video recordings, and students' written work (e.g., drawing of the student-made graph). Students were assigned pseudonyms for confidentiality. The interviews were transcribed, then all documentation and transcripts were coordinated in a software program for qualitative research. Initial analysis involved viewing the video recordings of all the interviews to generate a coding scheme for classifying the students' responses to the graphical prompts in the interviews. Next, two analysts independently coded the

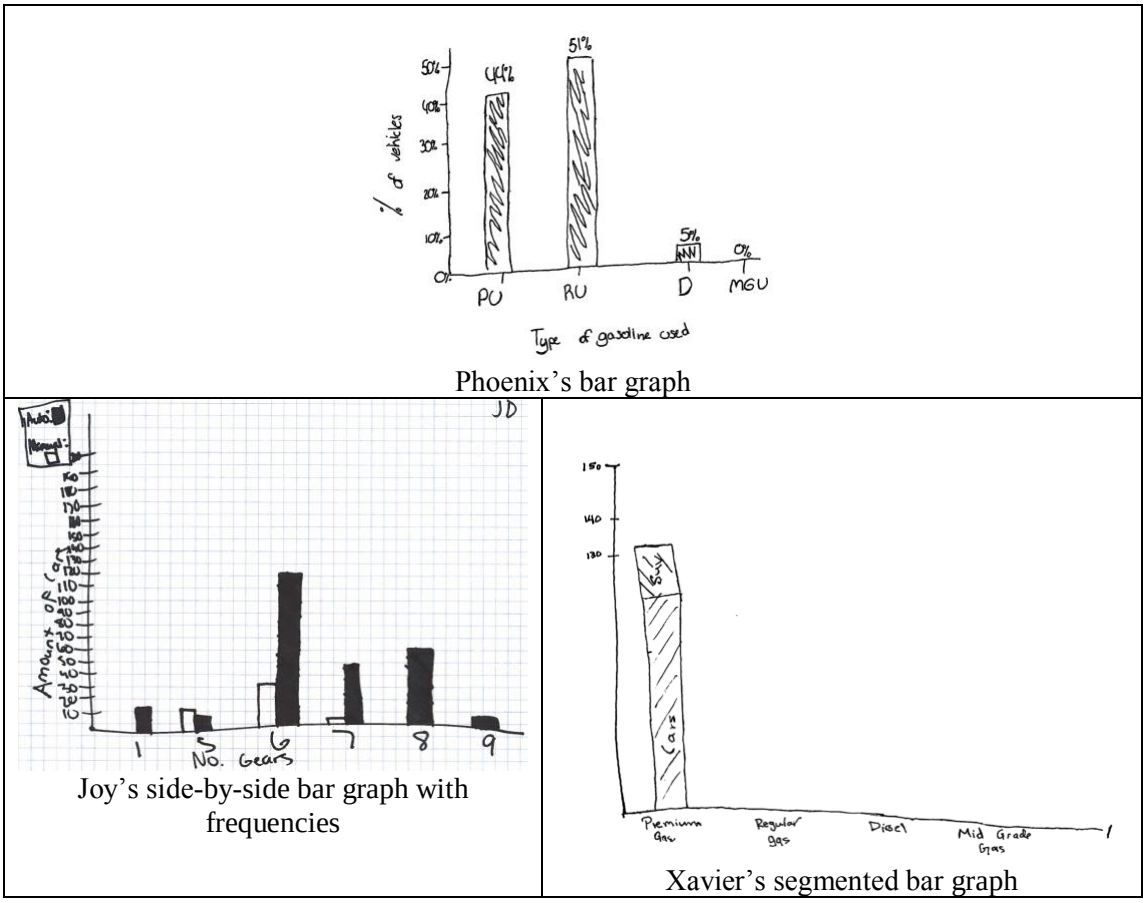
data through an iterative process, moving back and forth between documentation of the interviews and identifying, discussing, and classifying the responses. The student-made graphs were coded for type of graph, number of variables, whether the graph displayed frequencies or relative frequencies, and if one can assess association from the graph. Students' responses to the five graphs presented in part two of the interview were coded for the correctness of their interpretation of the graph and assessment of association of the variables displayed. Analysis of the coding resulted in the emergence of themes that are presented below.

RESULTS

Student-made Graphs

This subsection describes the graphs the students independently made after analyzing the Vehicle dataset. Recall that the students numerically (through two-way tables) investigated the relationship between two categorical variables of their choosing, then they were asked to sketch and interpret a graph that could be used to convince someone of the relationship they discovered.

Three of the thirteen students did not make a graph; Ella declined stating, "I don't know," and two other students re-created the two-way table. The remaining ten students' graphs were grouped based on whether one can assess association of the variables from the graph. Most of the students (six) created a graph from which you cannot assess association, as only one variable is displayed. This included three bar graphs, two pie graphs, and one box plot. Figure 2 presents Phoenix's bar graph displaying the type of gasoline used by all vehicles as an example from this group. Students in this group interpreted their graphs correctly, but either could not or incorrectly used their graphs to assess association when asked.



Phoenix's bar graph

Joy's side-by-side bar graph with frequencies

Figure 2. Student-made Graph Samples

Two of the students made graphs from which you can partially assess association, but because they used frequencies rather than relative frequencies it is not possible to do a thorough analysis from the graph alone. Kyle described transforming a two-way binned plot into a new graph

with a color gradient to indicate where points fell in the cells. "The ones [cells] that had more dots in them would be darker than the ones that didn't have them." Joy made a side-by-side bar graph that displayed the number of gears in vehicles, separated by type of transmission (automatic or manual); it is shown in Figure 2. Like the previous group of students, these two students correctly interpreted their graphs but incorrectly described how they could assess association using the graphs.

The remaining two students created graphs that make it possible to completely analyze whether the two categorical variables are associated. Xavier made a segmented bar graph (see Figure 2). In his graph, the height of a bar displays the number of vehicles that use that fuel type, and each bar's segmentation is based on the percent breakdown for the types of vehicles within that category. Note that on the graph, he has only drawn the bar for the 132 vehicles that use premium gas and segmented it to show that 87% of those vehicles are cars and the other 13% are SUVs. He verbally described completing the rest of the graph in the same manner, and properly described using the graph to convince someone of the relationship between types of vehicles and the fuel they use. The other graph in this group is Lana's side-by-side pie graph.

Students' Interpretations of Graphs for Bivariate Categorical Data

Table 1 identifies the percentage of the thirteen students who correctly and incorrectly interpreted each of the graphs presented in the second portion of the interview.

Table 1. Correctness of Students' Interpretations of Graphs

Graph	Correct Interpretation of Graph	Incorrect Interpretation of Graph	No Statement Made to Interpret Graph
Side-By-Side Bar Graph	46%	46%	8%
Segmented Bar Graph #1	62%	38%	0%
Segmented Bar Graph #2	92%	8%	0%
Eikosogram #1	31%	69%	0%
Eikosogram #2	46%	46%	8%

Students were generally most successful at interpreting the segmented bar graphs. 62% of students correctly interpreted the first segmented bar graph and 92% of students correctly interpreted the second one. Students were less successful at interpreting the other graphs, particularly the eikosograms. Only 31% of students interpreted the first eikosogram correctly and 46% of students interpreted the second eikosogram correctly.

Throughout the remainder of the results section, we share student responses regarding Segmented Bar Graph #1 that looked at the relationship between whether a bar has nuts and its texture. In their responses, three of the five students who incorrectly interpreted the graph misidentified the whole associated with a specific percent. For example, Hector discussed the graph as follows: "So this is saying that almost 70% of chewy bars don't have nuts." In his description, Hector made inferences about a relative frequency using an incorrect referent. Although the graph showed that roughly 70% of granola bars with no nuts or with nuts were chewy, Hector interpreted it as 70% of the chewy bars do not have nuts. In contrast, Oscar made an accurate interpretation of this graph with the correct referents:

> Only about 31% of the granola bars with no nuts were crunchy. The other 69% were chewy but had no nuts. Then the other one, 28% of granola bars with nuts were crunchy and 72% of granola bars with nuts were chewy. I guess.

Students' Interpretations of Association for Bivariate Categorical Data

Table 2 displays the results of analyzing the correctness of the students' statements regarding association of the displayed variables for each graph. Note that the last column shows that some students never made comments regarding the association of the variables, despite prompting.

Table 2. Correctness of Students' Interpretations of Graphs for Association

Graph	Correct Interpretation of Association	Incorrect Interpretation of Association	No Statement Made to Interpret Association
Side-By-Side Bar Graph (Association exists)	23%	31%	38%
Segmented Bar Graph #1 (No association exists)	23%	38%	38%
Segmented Bar Graph #2 (Association exists)	62%	31%	8%
Eikosogram #1 (Association exists)	23%	31%	46%
Eikosogram #2 (No association exists)	8%	77%	15%

For the side-by-side bar graph, 23% of students made correct interpretations, and 31% made incorrect interpretations. Students performed about the same on Segmented Bar Graph #1 but made many more correct statements noting association of the variables in the second segmented bar graph. For the two eikosograms, 23% and 8% of students correctly identified whether the variables were related in these graphs. It is important to point out that students were generally less successful at interpreting the graphs for association when no association between the two variables existed. The two graphs with no association had the highest percentages of incorrect interpretations of association by the students.

Students' responses regarding association between nuts and texture in granola bars based on analysis of Segmented Bar Graph #1 provide examples of how students reasoned about association. Andrea said the following:

> Because there is not really that much of a difference. Like there's more chewy bars for both categories and it's close to an even ratio. Because there is not really that much of a difference.

Cedric had a similar response, stating, "Not really, because they're around the same; there's not a drastic difference." These students' statements are representative of the three students who recognized that the two bars in the graph were segmented similarly and this similarity implied no association existed between the two variables.

In contrast, a larger number of students (5) incorrectly replied 'yes' when asked if a relationship existed between the two variables in Segmented Bar Graph #1. As an example, Joy stated the following:

> I just think that most granola bars overall are chewy and maybe so if there is more chewy bars. I mean if there is more with...this number [13] up here…. it should be… there are gonna be more chewy bars than crunchy bars.

Quela also stated that a relationship existed, saying "the chewy ones have more nuts than … the crunchy ones." These statements are representative of other students who incorrectly interpreted the graphs for association. Students struggled to know what to compare, and often compared frequencies rather than relative frequencies. In general, across the five graphs, only a few students correctly analyzed them to determine whether there was an association between the two variables under investigation. It is also important to point out that the number of students making correct claims about the variables' association across all five graphs was quite small. One student (Andrea) made correct statements about association for all five graphs, and another student (Cedric) made correct statements about four of the five graphs.

DISCUSSION

This study contributes to the ongoing research in statistics education about students' conceptions of statistical association by investigating students' reasoning about various types of graphs displaying bivariate categorical data. The students in this study had minimal prior

experience with representing or interpreting bivariate categorical data. Our study was designed to see how students interpret bivariate categorical data prior to instruction.

The results showed that most of the students struggled to create bivariate categorical data representations which could assess the association between the two variables. Many students created univariate graphs that represented only one of the variables; they were unable to create a display that coordinated the two variables. When they analyzed previously created graphs, a considerable number of students were able to interpret the three types of graphs considering this was prior to instruction. However, it was challenging for most students to determine from these graphs whether the two variables were associated to one another or not. Students commonly interpreted relative frequencies using the incorrect referent.

We also noticed that students' abilities to interpret a graph seemed to improve when they saw a graph type for the second time. A higher percentage of students interpreted the second segmented bar graph correctly than the first. A similar pattern was noticed for the eikosograms. Although these changes may be explained by students' increased familiarity with the graph type, we also believe that there may have been differences in the difficulty to interpret the graphs based on the data sets provided. Recall that the second segmented bar graph presented data in which one of the cells had zero occurrences; this may partially explain why students had fewer difficulties interpreting this graph. Furthermore, the first eikosogram was more challenging for students to interpret, perhaps because one of the variables had four different categories, and two of these categories included a small number of granola bars and only included one outcome for the other variable. Another possible explanation is that during the interview, the interviewer often questioned the students about the graphs in ways that may have impacted the students' thinking about the graphs. These questions and subsequent prodding may have influenced what students interpreted about that graph type later in the interview. Students' interpretation of association improved from the first to second segmented bar graph, but not for the eikosogram. We found that in both cases where no association existed (Segmented Bar Graph #1 and Eikosogram #2), students were hesitant to state that no relationship existed.

Students were most successful at interpreting the segmented bar graph, both as a graph and for association of the variables displayed. Although the eikosogram displays more information about the data set, students struggled to make sense of the graphs' meaning. Therefore, one of the curricular implications from our study is that when developing learning trajectories for categorical bivariate data, students should learn how to interpret segmented bar graphs prior to eikosograms. Likewise, in software programs utilized by students, the capability to make segmented bar graphs should be prioritized. Our future work will investigate how students and teachers use technology to investigate categorical data and create representations for bivariate categorical data.

REFERENCES

Australian Curriculum, Assessment and Reporting Authority (2013). *The Australian curriculum: Mathematics*. Sydney, Australia: Author.

Ball, D. L., Thames, M. H., and Phelps, G. (2008). Content Knowledge for Teaching: What Makes It Special? *Journal of Teacher Education, 59*(5), 389-407.

Batanero, C., Estepa, A., Godino, J.D., & Green, D.R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education, 27,* 151-169.

Bransford, J. D., Brown, A. L., and Cocking, R. R. (Eds.) (1999). *How people learn: Brain, Mind, Experience, and School.* Washington, D.C.: National Academy Press.

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In Batanero, C., Burrill, G., & Reading, C. (Eds.), *Teaching statistics in school mathematics - Challenges for teaching and teacher education. A Joint ICMI/IASE study: The 18th ICMI study* (pp. 57–69). New York, NY: Springer.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. New York, NY: Springer Science & Business Media.

Ministério da Educação (2017). *Base Nacional Comum Curricular*. Brasília, Brazil: Author.

National Governors Association Center for Best Practices & Council of Chief State School Officers (2010). *Common Core State Standards (Mathematics)*. Washington, DC: Authors.