FINDING WHAT FITS

Students explore six tasks to develop criteria for finding an informal line of best fit.

Stephanie A. Casey

Statistical association between two variables is one of the fundamental statistical ideas in school curricula (Burrill and Biehler 2011; Garfield and Ben-Zvi 2004). Indeed, reasoning about statistical association has been deemed one of the most important cognitive activities that humans perform (McKenzie and Middlesen 2007). Students are typically introduced to statistical association through the study of the line of best fit because it is a natural extension of their study of linear equations in mathematics. This is predominantly true for students in the United States; for example the authors of the Common Core State Standards for Mathematics (CCSSM) (CCSSI 2010) ask that students in eighth grade learn about linear equations, linear functions, and the line of best fit. A learning trajectory for linear regression study (Bargagliotti et al. 2012) begins with students finding and studying an informal line of best fit, which refers to the idea that students are fitting a line, by eye, to data displayed in a scatterplot, without making calculations or using technology to place the line. Hence, it is found informally. For example, CCSSM states that students should know the following:

> Know that straight lines are widely used to model relationships between two quantitative variables. For scatter

CHUNG LEE

> plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line. (p. 56)

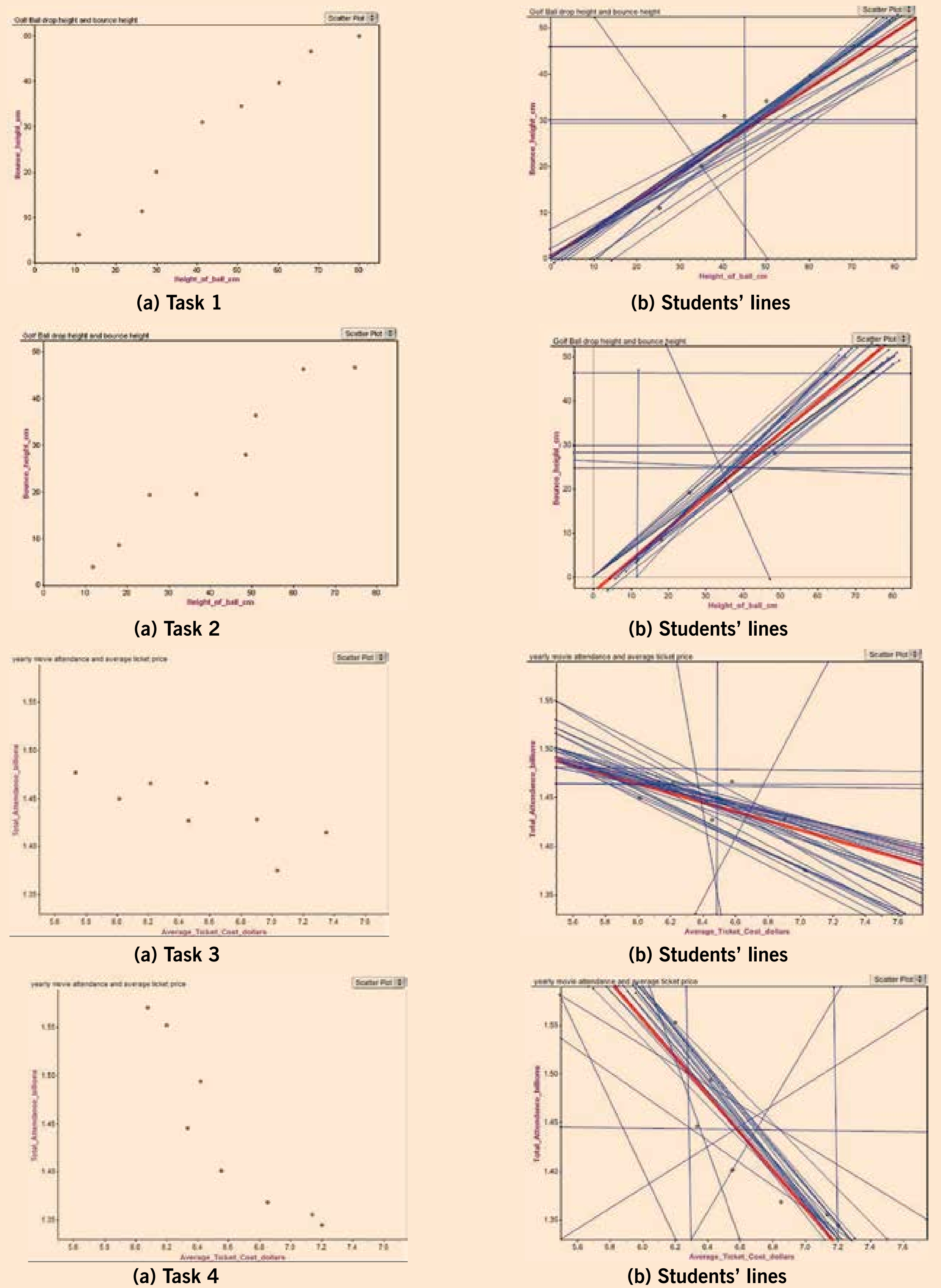
The Common Core Standards Writing Team (2011) specified that this standard includes an expectation that students determine that the informal line of best fit for data that has no association should be a horizontal line, and that a horizontal fitted line implies that there is no association between the variables.

This article shares responses to a series of six tasks from a study analyzing students' understanding of the informal line of best fit. Thirty-three eighth-grade students in the United States were interviewed before they received instruction on the line of best fit (Casey 2015). Teachers can benefit from learning about this study in multiple ways. They can acquire meaningful tasks to implement with students when teaching informal line of best fit; gain knowledge of conceptions that students have about the line of best fit to plan for and manage instruction on the topic; and learn other implications for teaching the topic that resulted from the study.

DESCRIPTION OF TASKS

The first five tasks asked students to place a piece of piano wire to represent the line of best fit for data presented in a scatterplot and justify why they placed it there. Piano wire was chosen for its rigidity and thinness, although in other settings the tasks have been completed equally well using raw spaghetti or pipe cleaners. The five tasks implemented are displayed in **figures 1** and **2**. The data were chosen on purpose. The plots (1) presented data from real-world contexts that were familiar to students; (2) had eight points, which was a manageable number; and (3) did not contain outliers or influential points. They progressed from plots displaying a strong positive association (tasks 1 and 2), to plots displaying a relatively strong negative association (tasks 3 and 4), to a plot displaying no association (task 5).

Fig. 1 Scatterplots were presented to students (a), who then placed best-fit lines (b). The least-squares regression line is in red.



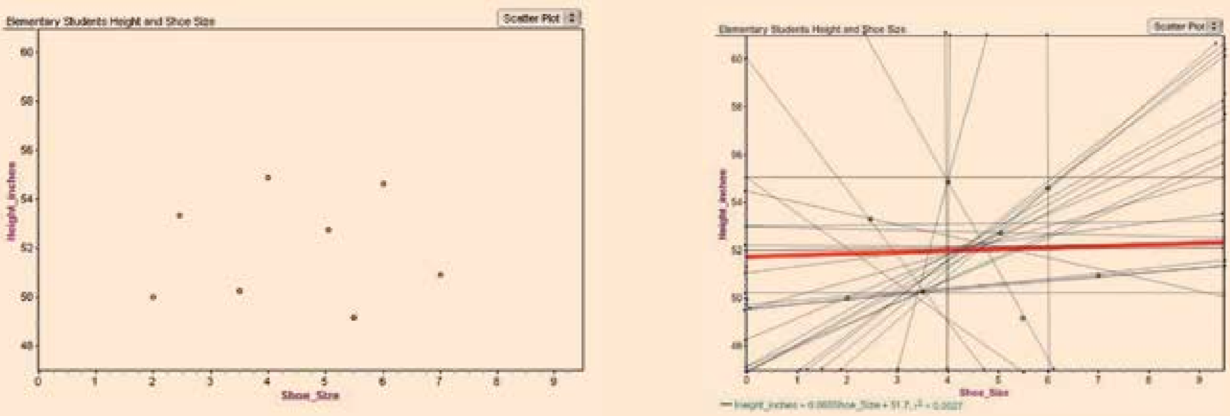
RESULTS: THE LINE OF BEST FIT WITH LINEARLY ASSOCIATED DATA

The first notable result was that a sizeable number of students (9), when asked to find the line of best fit on the first task, wanted to bend the wire to connect the points on the scatterplot. For instance, Marcus (a pseudonym, as are all student names) asked, “Wouldn’t it be like the line that starts here [the origin] and like, connects . . . connects all these points, right?” Some students struggled to conceive of the line of best fit as a line that did not necessarily go through all the points, likely because this differed from graphs of linear functions that these students had been studying in mathematics. When statements like this occurred after students were presented with the first task, the interviewer redirected by explaining that the goal was to find the line of best fit. Because lines are straight, students were not to bend the wire. After receiving this instruction, all the students were able to complete the tasks, suggesting that this same redirection may be effective in a classroom setting.

Figure 1 presents all 33 students’ lines for tasks 1 through 4. The least-squares regression line plotted in red provides a visual image of the accuracy of the placed lines for these tasks. These displays show that there was considerable variability in the placed lines’ locations. The majority of the lines were reasonably accurate in that they were generally close to the least-squares regression line, but a substantial number of lines were placed inaccurately. Looking at the criteria that students used for placing the lines provided more insight into the process (see **table 1** for the students’ criteria and the number of different students who used each criterion).

Table 1 reveals that the criteria that students naturally devised for finding the informal line of best fit were numerous and varied in their viewing of the data set as a whole.

Fig. 2 Task 5 explored elementary students’ height and shoe size.



(a) Task 5 **(b) Students’ lines**

Equal number points each side (7)
Closest to all the points (6)
Most points (4)
Through first point and last point (4)
Context (4)
Horizontal through highest point
Horizontal halfway between highest and lowest points
Horizontal at low value
Through middle point
Through maximum point
Through bottom-left corner and direction
Connect point on *y*-axis of highest point’s *y*-coordinate to point on *x*-axis of *x*-coordinate of lowest point

(c) Students’ approaches

Table 1 The description and number of students who used various criteria showed contrasting views.

Criteria	Number of Students
Through as many points as possible	13
Equal number of points on both sides	8
As close to all the points as possible	7
Reflect the relationship the variables have on the basis of context knowledge	5
Halfway between the lowest and highest points	3
Through the first and last points	3
Starting from the first point then maximizing the number of points it goes through	2

Some criteria used the selection of specific points (e.g., lowest and highest, first and last) to determine the line, ignoring the rest of the data set. Other criteria, such as "equal number of points on both sides" and "as close to all the points as possible," showed that the students were considering the data in their entirety when finding the line of best fit. The third most commonly used criterion, "as close to all the points as possible," is the one encouraged by CCSSM (CCSSI 2010) and is in agreement with the approach of the least-square regression line.

A closer examination of the criteria for and the location of lines placed on task 2 provided greater insight regarding students' conceptions of the line of best fit. **Figure 1**, task 2 (b) shows the informal best-fit lines that students placed on task 2 along with the least-squares regression line. The thirteen criteria identified by the 33 students when placing the line on this task (see **fig. 3**) resulted in a large number of lines placed near the least-squares regression line. However, most generally ran parallel to or split the least-squares regression line, with very few following it. This occurred because of the predominance of the most points and equal number criteria and the decision of students employing those criteria to force their line to go through one of the last two points.

A closer examination of the lines placed by students so that an equal number of points would be on each side of the line (see **fig. 4**) revealed that this criterion resulted in remarkably different lines. Three of these lines were relatively accurate, with one following the least-squares regression line nearly exactly. However, the other two lines were inaccurate because they were placed horizontally. These students' explanations about the horizontal placement sound appropriate ("I'm putting it in the middle"; "It's at the average"), and a teacher would be inclined to think that these students understood the topic. However, these students applied "middle" and "average" in an univariate rather than bivariate sense and therefore placed their lines at the "middle" or "average" of the bounce height only.

These explanations and actions should raise cautions for teachers when teaching the topic: avoid solely teaching students to place the line so that an equal number of points are on each side and probe what your students mean by "middle" and "average" in a bivariate data analysis setting.

RESULTS: THE LINE OF BEST FIT FOR DATA WITHOUT ASSOCIATION

The presentation of task 5's scatter plot that displayed no association evoked different responses and approaches from the students than the previous four tasks (see **fig. 2**). The time it took students to complete this task was considerably longer than the other tasks, and many students studied the plot in silence for a substantial time (around twenty seconds) before responding. Six students initially commented that they did not see a general trend or direction in the plot and were confused about what to do. One student, however, commented that she did not see a general trend in the plot but correctly used that observation to place the line both horizontally and halfway between the lowest and highest points because "it's not decreasing or increasing." This is the conclusion we wanted to help all students make (Common Core Standards Writing Team 2011), but it was evidently not a natural conclusion for students.

There were various locations for the placed lines on this task compared with the previous four tasks. **Figure 2b** displays all the lines placed by the students (Sasha said, "I have no idea," and did not place a line), along with the least-squares regression line. The criteria employed by students on this task ordered by frequency of use are listed in **figure 2c**. The number of students choosing a criterion was shown in parentheses if used by multiple students.

It is notable that relatively few students placed lines close to the least-squares regression line. Even those students who claimed to place the line closest to all the points, as the least-squares regression essentially did, were unable to do so accurately on this task. Another important observation to make from **figure 2** is that a large number of the placed lines have positive slopes likely because students expected that bigger shoe sizes correlated to bigger heights. Therefore, they placed their lines with positive slopes to show

Fig. 3 Students used several criteria when placing the line for task 2.

Most points
Through (0, 0) and last point
Through (0, 0) and most points
Through first point and last point

Equal number points each side
Through a point and equal number points each side

Closest to all the points
Through (0, 0) and closest to all the points
Through first point and last point
Horizontal through highest point
Horizontal at average of *y*-variable
Through middle point, intersect middle of *x*-axis
Vertical to hit *x*-axis at value of lowest point, upper end across from highest point

Reasoning about statistical association has been deemed one of the most important cognitive activities that humans perform.

that relationship although it was not exhibited in the data in the plot. One teaching implication is that students should be asked to work with data sets such as this one that disagree with assumed relationships to encourage students to discuss what to base the placement of the line of best fit on: contextual knowledge, the data at hand, or some combination of the two.

EVALUATING LINES OF BEST FIT

A classroom of students informally fitting a line of best fit to data will result in numerous lines, so it is important that students consider how to evaluate lines to determine which line best fits the data. To this end, a sixth task was presented to students in the study. The scenario for this task was that two students, Angelo and Barbara, were asked to complete task 1 but had different solutions (see **fig. 5**). Students were asked, "Which student's line fits the data better and why?" The task was designed so that Angelo and Barbara's line placement would be similar; however, Angelo's line (A) goes through two points, whereas Barbara's line (B) was closest to all the points (it was the least-squares regression line) but did not go through any points.

One-third (11) of the students in the study chose line A; the other two-thirds (22) chose line B. Seven of the 11 students who chose line A stated that they preferred it because it went through some of the points, including 3 students whose dominant criterion for placing lines was through the most points. Thus, teachers can anticipate that a sizeable number of their students will likely need learning experiences to change their conception that it is more important to go through, rather than be near, all points (the criteria included in CCSSM 8.SP.A.2; CCSSI 2010).

Nineteen students who chose line B explained that it was closer to all the points than line A. One notable result was that 7 of the 10 students whose dominant criterion for placing the line of best fit on tasks 1–5 was "through the most points" chose line B as the better line, shifting to note that being closest to all the points was most important for the line of best fit. For 3 of these students, their progression through the tasks involved a transition away from the criteria of "through the most points" that they had used for the earlier tasks.

As Sasha described, she "started out thinking like Angelo but now sees

Fig. 4 Students placed lines with an equal number of points on each side.

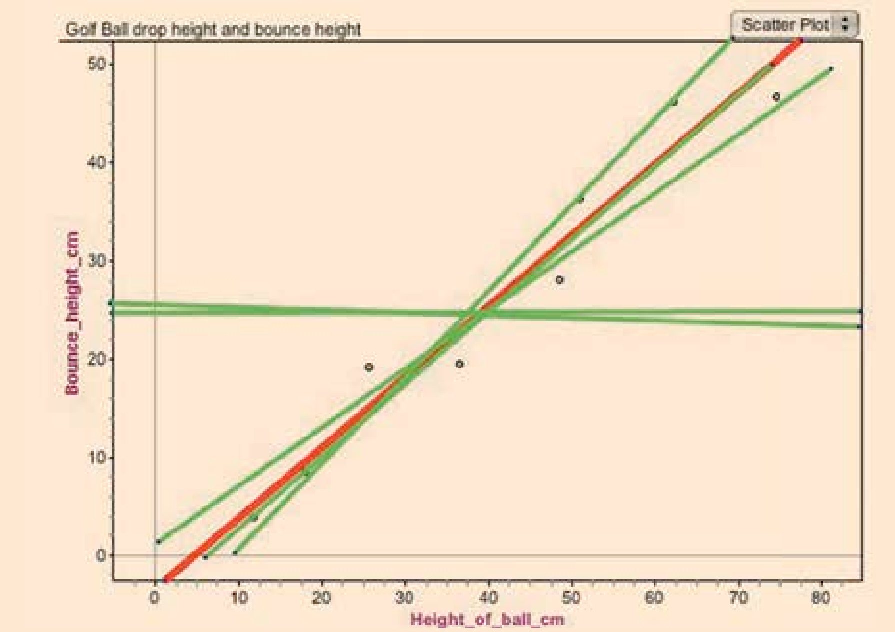
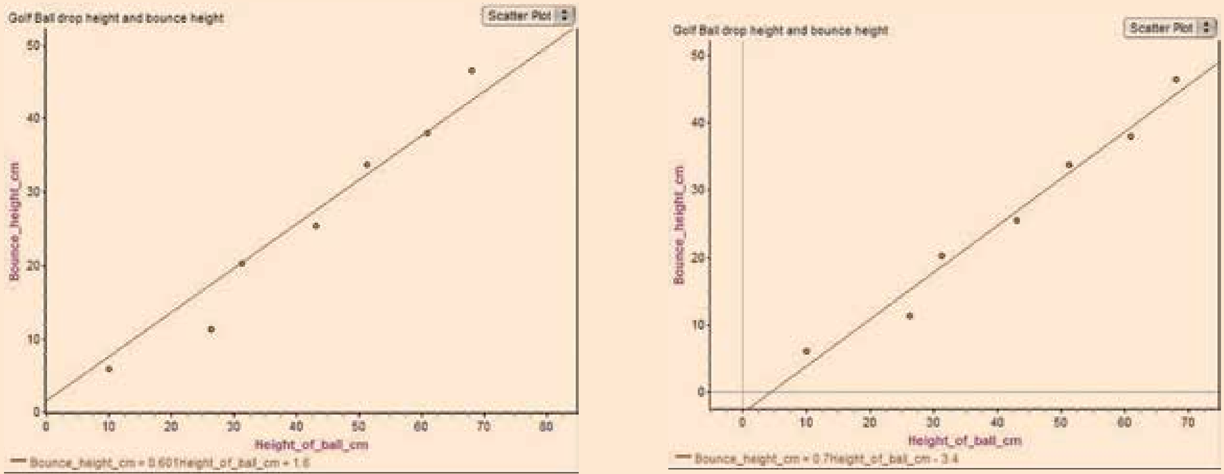


Fig. 5 Task 6 from the interview asked students, "Which line fits the data better and why?"



Line A

Line B

that Barbara's is better." For others, completing this task was an illuminating experience. It allowed them to evaluate whether going through or being near all the points was more important. For a number of students, that evaluation process helped them see why being closer to all the points created a better line of best fit. Teachers are encouraged to use this task for those same purposes in their classrooms.

MEANINGFUL IDEAS AND ESSENTIAL KNOWLEDGE

The informal line of best fit is a relatively new addition to the mathematics curriculum with the implementation of CCSSM (CCSSI 2010); however, it is extremely important because it serves as the foundational topic for the study of the fundamental concept of statistical association. The tasks and student responses to them described how students conceive of the informal line of best fit. In so doing, instruction might be crafted to meet students' learning needs.

ACKNOWLEDGMENTS

The author wishes to thank David Wilson for his collaborative work on this study. For more information, read the published lesson plan called "What Fits?" (Bargagliotti and Casey 2013) in the American Statistical Association's Statistics Education Web (STEW), which is based on the same study and contains additional tasks that teachers can use to teach the topic.

REFERENCES

Bargagliotti, Anna, Celia Anderson, Stephanie Casey, Michelle Everson, Chris Franklin, Rob Gould, Randall Groth, John Haddock, and Ann Watkins. 2012. "Project-SET Linear Regression Learning Trajectory June 2014." Project Set. http://

projectsetdotcom.files.wordpress.com/2014/06/regression-lt-final.pdf
Bargagliotti, Anna, and Stephanie Casey. 2013. "What Fits?" Statistics Education Web. http://www.amstat.org/education/stew/
Burrill, Gail, and Rolf Biehler. 2011. "Fundamental Statistical Ideas in the School Curriculum and in Training Teachers." In *Teaching Statistics in School Mathematics—Challenges for Teaching and Teacher Education*, *A Joint ICMI/IASE Study: The 18th ICMI Study*, edited by Carmen Batanero, Gail Burrill, and Chris Reading, pp. 57–69. New York: Springer.
Casey, Stephanie. 2015. "Examining Student Conceptions of Covariation: A Focus on the Line of Best Fit." *Journal of Statistics Education* 23 (1). http://www.amstat.org/publications/jse/v23n1/casey.pdf
Common Core State Standards Initiative (CCSSI). 2010. Common Core State Standards for Mathematics. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers. http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf
Common Core Standards Writing Team. 2011. *Progressions for the Common Core State Standards in Mathematics* (Draft). http://commoncoretools.files.wordpress.com/2011/12/ccss_progression_sp_68_2011_12_26_bis.pdf
Garfield, Joan, and Dani Ben-Zvi. 2004. "Research on Statistical Literacy, Reasoning, and Thinking: Issues, Challenges, and Implications." In *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking*, edited by Dani Ben-Zvi and Joan Garfield, pp. 397–409. Dordrecht, The Netherlands: Kluwer Academic Publishers.
McKenzie, Craig RM, and Laurie A. Middlesen. 2007. "A Bayesian View of Covariation Assessment." *Cognitive Psychology* 54 (1): 33–61.

CCSSM Practices in Action

8.SP.A.1–2
8.EE.B.5–8
8.F.A.2–4

Stephanie A. Casey, scasey1@emich.edu, is a mathematics teacher educator at Eastern Michigan University in Ypsilanti. She is interested in the teaching and learning of statistics at the middle and secondary levels, motivated by her experience of teaching secondary mathematics for fourteen years.