When Data and Chance Collide:
Drawing Inferences from Empirical Data

Chapter to appear in
National Council of Teachers of Mathematics 2006 Yearbook
Thinking and Reasoning with Data and Chance
Edited by Gail Burrill

James E. Tarr
Department of Learning, Teaching, & Curriculum
303 Townsend Hall
University of Missouri
Columbia, MO 65211-2400
Work Phone: 573.882.4034
E-mail: tarrj@missouri.edu

Hollylynne Stohl Lee
Department of Mathematics, Science, & Technology Education
326 Poe Hall – Campus Box 7801
North Carolina State University
Raleigh, NC 27695-7801
Work Phone: 919.513.3544
E-mail: hollylynne@ncsu.edu

Robin L. Rider
Department of Mathematics & Science Education
321 Flanagan
East Carolina State University
Greenville, NC 27858-4353
Work Phone: 252.328.9369
E-mail: riderr@mail.ecu.edu

When Data and Chance Collide:
Drawing Inferences from Empirical Data

Ms. Rodriguez circulated among her sixth-grade students, listening carefully to their reasoning about data. Their task was to collect "compelling evidence" in order to decide whether a die is fair or biased. She was pleased by the mathematical discourse that was elicited as pairs of students rolled a "virtual die," collecting empirical data, and monitoring the frequency and relative frequency of each outcome, 1-6:

Brandon: [After 588 trials] I really don't think it's fair.

Manuel: Every single thing doesn't have to be even, man. It's just the luck. They are pretty much close.

Brandon: I still think that 5 is continuously behind [in the bar graph].

Manuel: If you don't think this is fair… It's fair, man.

Brandon: But look at the 5.

Manuel: It doesn't all have to be perfect, man!

Brandon: [At 650 trials] Look at the percents: 13% for the 5, 13% for the 3, 20% for the 2 and the 4… I bet you it's weighted. I bet you we aren't fair.

Manuel: I bet you're wrong. I bet we are fair. Just because it's not all even doesn't mean we're not fair.

The vignette above focuses on differences in students' attention to variation in data as they begin to build informal notions of statistical inference. It also reveals the kinds of diversity in students' reasoning that teachers encounter in teaching probability and statistics, and raises several key questions:

- What do students consider "compelling evidence" in formulating and evaluating arguments based on data?

- To what extent are students aware of the role sample size plays in making inferences from data?
- How much variation do students consider “tolerable” between their predictions (or what is expected) and actual outcomes?

Principles and Standards for School Mathematics (National Council of Teachers of Mathematics [NCTM] 2000) articulates the importance of reasoning about and from data for all students, and for Grades 6-8 advocates, “Upper-elementary and early middle grades students can begin to develop notions about statistical inference” (p. 50). The purpose of this chapter is to share the insights we gained from implementing a task with sixth-grade students as they learned to draw inferences from empirical data. To accomplish this goal we begin by describing the key features of the task that elicit and extend students’ reasoning. Next we provide several contrasting examples that exemplify the notion of “compelling evidence” among middle grades students, and then offer provisions for individual differences. Finally we argue that carefully-designed instructional tasks can engage students of all different ages in statistical inference and promote the development of powerful connections between data and chance.

Problem Tasks that Promote Connections Between Data and Chance

Elementary and middle school students hold strong beliefs regarding the fairness of dice and, in fact, generally doubt that each outcome is equally probable (Watson and Moritz 2003). Such pervasive beliefs are likely a product of game-playing experiences and represent genuine challenges to mathematics teachers. The task we designed sought to capitalize on students’ experiences and attention to “fairness” by engaging them in data collection and analysis to conjecture whether die from a particular company is fair or biased. In the *Schoolopoly* task (see fig. 1), students utilize dynamic computer software, *Probability Explorer* (Stohl 1999-2004) or the *ProbSim* application on TI-73, TI-83+ and TI-84+ graphing calculators to roll a “virtual die” and generate large samples quickly. For each die company (see fig. 2), weights are assigned to each event, 1-6 and can remain hidden from students during the activity. Moreover, empirical data can be represented in a variety of displays using both technology tools (see fig. 3).

Schoolopoly: Is the die fair or biased?

Background
Suppose your school is planning to create a board game modeled on the classic game of *Monopoly*. The game is to be called *Schoolopoly* and, like *Monopoly*, will be played with dice. Because many copies of the game expect to be sold, companies are competing for the contract to supply dice for *Schoolopoly*. Some companies have been accused of making poor quality dice and these are to be avoided since players must believe the dice they are using are actually “fair.” Each company has provided dice for analysis and you will be assigned one company to investigate:

Luckytown Dice Company
Dice R’ Us
High Rollers, Inc.
Dice, Dice, Baby!
Pips and Dots
Slice ‘n’ Dice

Your Assignment
Working with a partner, investigate whether the dice sent to you by the company is *fair* or *biased*. That is, collect data to infer whether all six outcomes are equally likely and answer the following questions:

1. Do you believe the dice you tested are fair or biased? Would you recommend that dice be purchased from the company you investigated?
2. What *compelling evidence* do you have that the dice you tested are fair or unfair?
3. Use your data to estimate the probability of each outcome, 1-6, of the dice you tested.

Collect data about the dice supplied to you. Note that each single trial represents the outcome of one roll of a “new” virtual die provided by the company.

Copy any graphs and screen shots you want to use as evidence and print them for your poster. Give a presentation pointing out the highlights of your group’s poster.

Fig. 1. *Schoolopoly* task

Company	Weight [P(1)]	Weight [P(2)]	Weight [P(3)]	Weight [P(4)]	Weight [P(5)]	Weight [P(6)]
Luckytown Dice Company	3 *[0.15]*	3 *[0.15]*	3 *[0.15]*	3 *[0.15]*	3 *[0.15]*	5 *[0.25]*
Dice R' Us	2 *[0.125]*	3 *[0.1875]*	3 *[0.1875]*	3 *[0.1875]*	3 *[0.1875]*	2 *[0.125]*
High Rollers, Inc.	2 *[0.1333]*	3 *[0.2]*	2 *[0.1333]*	3 *[0.2]*	2 *[0.1333]*	3 *[0.2]*
Dice, Dice, Baby!	2 *[0.1111]*	3 *[0.1667]*	4 *[0.2222]*	4 *[0.2222]*	3 *[0.1667]*	2 *[0.1111]*
Pips and Dots	1 *[0.1667]*	1 *[0.1667]*	1 *[0.1667]*	1 *[0.1667]*	1 *[0.1667]*	1 *[0.1667]*
Slice 'n' Dice	4 *[0.16]*	5 *[0.2]*	5 *[0.2]*	5 *[0.2]*	1 *[0.04]*	5 *[0.2]*

Fig. 2. Weights [and *corresponding theoretical probabilities*] for each event in *Schoolopoly* dice companies

Alternatively, teachers might have students approach the problem using hands-on materials such as foam dice (sold in teacher supply stores) or weighted dice (sold in novelty stores); web sites even provide instructions for creating a 'homemade" biased die by folding paper to form a cube. In particular, students can explore whether these alternative dice are essentially equivalent to standard number cubes, or whether, in fact, there exists empirical evidence to support the notion that these dice are biased. The use of hands-on materials makes the *Schoolopoly* task more accessible while maintaining opportunities for students to engage in data-based decision-making. However, this alternative approach is less efficient because of the additional time required to generate large samples of data and representations of data that serve as the basis of subsequent student inferences.

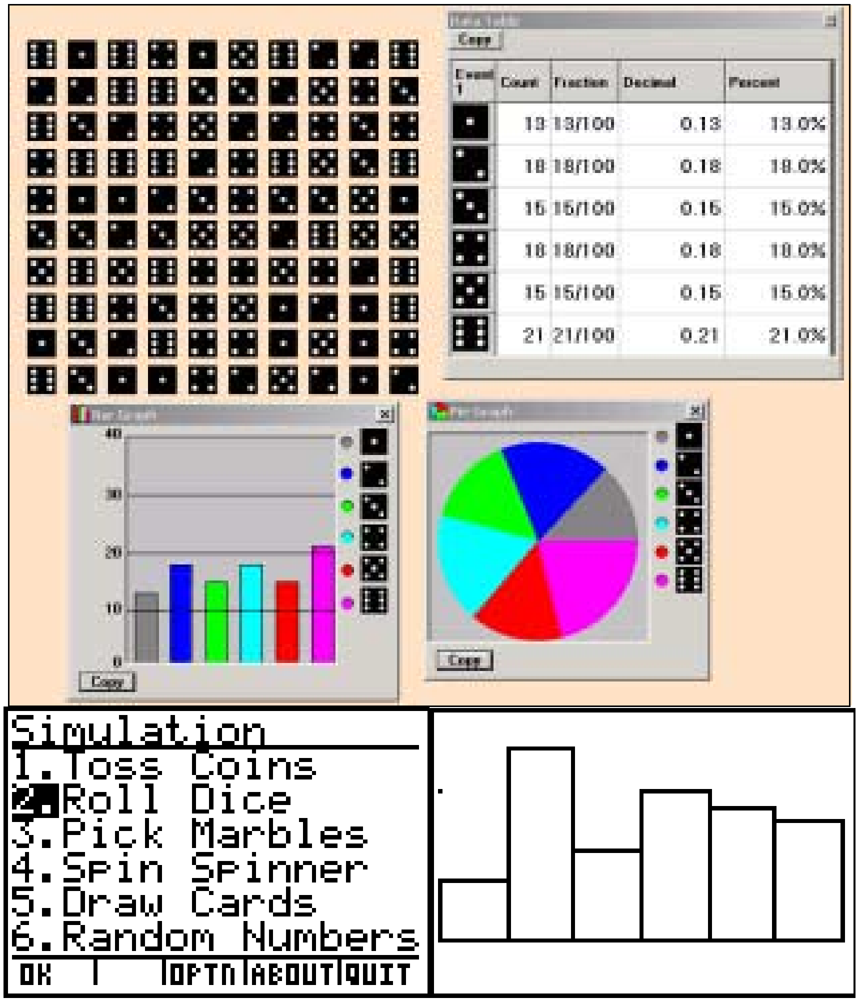

Fig. 3. Screen shots depicting results of 100 rolls of "virtual dice" as shown in *Probability Explorer* and a TI-83+ graphing calculator

Key Features of the Task

In addition to the captivating problem context, problem tasks should embody the tenets of the Data Analysis & Probability Standard (NCTM 2000) by requiring students to "formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them;" "select and use appropriate statistical methods to analyze data;" and "develop and evaluate inferences and predictions that are based on data" (p. 48).

In the *Schoolopoly* task, students are required to both make a decision regarding the fairness of a die and also to support that decision with compelling evidence rather than mere opinions.

Moreover, in formulating data-based arguments, pairs of students must negotiate what constitutes “compelling evidence,” use their argument to convince their partner, and ultimately defend their decision to peers in a whole-class setting.

A critical idea in understanding statistics is the notion that larger samples yield more power in making inferences, thus affording more confidence in the validity of one’s conclusions. Yet research indicates students are often unaware of the role sample size plays in drawing inferences (e.g., Aspinwall and Tarr 2001), and school mathematics curricular materials typically do address this fundamental principle. Our problem task seeks to foster the development of this idea by purposefully not prescribing a sample size. Consequently, students possess control in determining how many trials are sufficient to support their conclusions. Thus, arguments based on smaller samples become vulnerable to the scrutiny of classmates, whose inferences are grounded in larger samples, calling into question what constitutes “compelling evidence.”

Another key feature of the task is that theoretical probabilities are not determined from symmetry of the die, numerical computation, geometrical measures, nor by simply examining the die. Instead, students must *collect data* and *reason from data* in order to estimate (unknown) theoretical probabilities. There are many real-world situations for which the probability of an event can be estimated only through data collection. For example, the proportion of adults who smoke cigarettes is estimated only by collecting data from the general population. In this problem task, students engage in statistical inference and employ related principles such as the Law of Large Numbers which states that, for a given event, the empirical probability is more likely (although not certain) to approximate the theoretical probability as sample size increases.

Finally, students must decide how data should be displayed in formulating their own conclusions as well as communicating their arguments to classmates. In particular, data displays such as pie graphs are particularly useful in making part-whole comparisons while bar graphs assist students in making part-part comparisons. Tools like *Probability Explorer* and graphing calculator features such as the *ProbSim* application aid in this process by allowing students to choose an appropriate sample size and efficiently collect large amounts of data and display data in a variety of ways. In addition, both types of technology tools can update displays of empirical results after each trial, freeing up student resources to focus on relevant features of displays and attend to variation *during* and *after* the data collection process.

Learning to Draw Inferences:

Student Conceptions of "Compelling Evidence"

An important aspect of formulating and evaluating inferences is an understanding of the *unpredictability* of random phenomenon in the short-run but *predictability* in the long-run trends in data, that is, data from larger samples; the two different data sets from 100 rolls of a fair die (fig. 3) demonstrate the wide variability that is possible in small samples. Recent research (e.g., Aspinwall and Tarr 2001; Metz 1999; Stohl and Tarr 2002) on instruction has shown that developing intuitive notions about sample size fosters student understanding about the power of larger samples in making better inferences. Whether rolling "virtual dice," weighted dice, or standard number cubes, students engaged in the *Schoolopoly* task must grapple with questions such as:

- How many trials of the experiment should we carry out?
- Should we pool (combine) our empirical data with our previous results or keep our results independent of previously-collected data?
- Should we compare empirical data to our predictions (expectations) or compare between outcomes, 1-6?

Interestingly, students in Ms. Rodriguez's class offered a variety of strategies to address the notion of compelling evidence. For example, as data were collected, many students carefully monitored results by observing fluctuations to sectors of the pie graph and in the relative frequencies within the data table. These students ceased data collection and rendered decisions regarding fairness only after relative frequencies became more stable (after 300 or more trials). In contrast, others used smaller sample sizes but carried out many independent sets of trials before searching for consistencies in the results. Thus, students used "5 came in last every time" as a viable argument for justifying the biased nature of the die. Still others put unwarranted confidence in small samples of empirical data and were challenged by classmates when presenting recommendations in subsequent whole-class discussions.

Differences in student conceptions are further evident in the posters of figure 4 and figure 5, each depicting a conclusion regarding dice produced by the Dice R' Us company (fig. 2). In particular, Dannie and Lara (fig. 4) presented results from 36 trials as evidence that their company was worthy of a contract because variation between outcomes was acceptable to them. The class, serving as the "school board" to evaluate inferences, did not place much faith in Dannie and Lara's argument because of the small number of trials they used. Interestingly, students questioned this faulty claim even before realizing that the next group of presenters (Maria and Taiesha) had independently collected much more data about the same company.

Fig. 4. Dannie and Lara's poster depicting inferences based on a small sample.

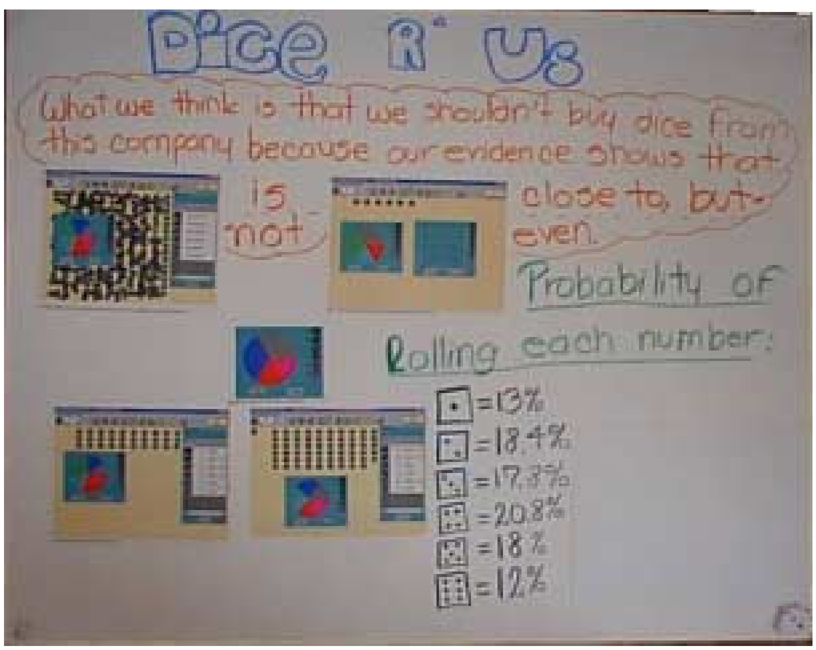

Fig. 5. Maria and Taiesha's poster depicting inferences based on larger samples.

In contrast, Maria and Taiesha (fig. 5) carried out nearly 1000 trials and argued against offering a contract to their company. In whole-class discussions, several classmates compared the two posters and asked each group to defend why they had chosen their particular sample sizes. After closely examining both posters, members of the “school board” voted in secret ballot, and 23 of 24 students judged the evidence presented in figure 5 as more compelling and decided that Dice R’ Us likely produced unfair dice; thus, at least one of either Dannie or Lara was more convinced by the second group’s evidence and voted accordingly. It is important to note that although voting is not an appropriate method for determining what constitutes “compelling evidence,” it nevertheless served as an efficient means of assessing students’ beliefs about statistical inference.

A related factor in determining which evidence is “compelling” is how much variation is “tolerable” among outcomes. In particular, Manuel’s remark from the opening vignette that “everything doesn’t have to be even… it’s just the luck” reflects his willingness to accept variability within a reasonably large sample of empirical data (n = 650). In contrast, Brandon argued that differences within empirical data were, in essence, significantly large enough to question the notion of fairness. Both Manuel and Brandon’s perspectives represent key principles in data analysis: (a) repeated experiments yield a variety of results due to randomness and variability that should be expected, and (b) “discrepancies between predictions and outcomes from a large and representative sample must be taken seriously” (NCTM 2000, p. 253). Students must coordinate these two seemingly conflicting ideas when drawing inferences. Students engaged in our task confronted the discrepancies between relative frequencies and questioned whether they were, in fact, large enough to be taken seriously.

Increasing Accessibility: Provisions for Equity

Teachers regularly encounter a wide-range of sophistication in students' statistical reasoning and "equity requires accommodating differences to help everyone learn mathematics" (NCTM 2000, p. 13). Problem tasks such as *Schoolopoly* can be modified to address such diversity by providing different levels in which to engage students, varying the problem sophistication, and even changing the problem's intended audience.

Level of Student Engagement

"Assessment should be an integral part of instruction that informs and guides teachers as they make instructional decisions" (NCTM 2000, p. 22) and tasks such as *Schoolopoly* provide opportunities for intertwining assessment and instruction. Ideally, problem tasks should elicit particular conceptions and misconceptions, enabling students to reflect on the validity of statistical arguments, and providing teachers access to students' reasoning with data and chance.

We found that this task effectively provided a window into the wide range of student reasoning and thus served as a useful foundation for subsequent discussions. Moreover, teachers can tailor their questioning based on the sophistication of student reasoning. For example, students who seem unaware of the role of sample size in drawing inferences could be asked, "Why not carry out only 6 trials of the experiment?" and "Is it possible to roll the die 6 times and never get a 3?" These questions seek to develop the notion that sample size *does* matter because small samples often lead to erroneous conclusions such as "It is impossible to roll a 3." Students who base inferences on relatively small samples (e.g., 50 trials) could be asked, "If you were to carry out the experiment again, would you expect the same results? If the results were different, which data set would you use as the basis of your conclusions? Why?" In contrast, students who realize the power of larger samples could be asked, "When did you notice the relative frequencies ‘evening out’? If they stabilized before 500 trials, what are the advantages for carrying out additional trials?” Finally, students could be asked to quantify their estimates of confidence in their decision by asking, “How confident are you after 100 trials? What about after 500 trials, or 1000 trials? Can you use numbers to describe your level of confidence?” Answers to questions such as these give rise to discussions of opinion polls that typically report measures of error in results (e.g., “± 4%”) and lay the conceptual foundation for higher-level processes of statistical inference.

Level of Problem Sophistication

The *Schoolopoly* task can be used with a variety of students, middle school through college-level, by modifying the learning expectations. For example, our study shows the task to be effective in developing the notion of statistical inference among middle school students but the same problem can be used with high school and college-aged students to stimulate debate about what constitutes “compelling evidence” when making inferences. Such rich discussions serve to promote the fundamental principle that sample size plays a critical role in statistical inference. Furthermore, teachers of Advanced Placement (AP) Statistics or college-level courses can engage students in more sophisticated analyses by having students use empirical data as the basis of a formal application of inferential statistics, applying use of a Chi-Square Test for Goodness of Fit to determine if differences in the observed six proportions of outcomes are indeed statistically significant from the expected 16.67% on a fair die. Typically, these students experience inferential statistics strictly from a theoretical viewpoint. These activities promote connections between empirical data collection and theoretical computation of test statistics.

High school and college students can also determine what sample size is sufficiently large. In observing the relative frequencies of each outcome as the sample size grows incrementally, students begin to notice the "evening out" of sectors of the pie graph or percentages in the table. At what point do the empirical probabilities become relatively stable? In other words, if 2000 trials is powerful enough to conclude the die is weighted, is 1000 trials sufficient? What about 500 trials? These questions can be used to motivate building confidence intervals around an expected proportion and studying the effect on the margin of error and interval width as sample size increases. Additionally, students can explore and discover that the "stabilization point" for relative frequencies is not the same for all problems. In the *Schoolopoly* task, more rolls are required for relative frequencies to stabilize for companies such as Slice 'n' Dice that include a rare outcome; for a more evenly distributed die, fewer rolls are required for relative frequencies to stabilize.

Likewise, the problem can be modified for younger audiences who can focus on the relationship between empirical and theoretical probability. By carrying out smaller sets of trials, they can learn that small samples are more likely to produce "unusual" or unrepresentative results such as no 3s in 10 tosses of the die. They can also learn that no finite number of rolls guarantees the occurrence of any outcome, 1-6, and explore concepts such independence.

In short, tasks such as *Schoolopoly* afford all students access to a real-word problem that requires them to collect and analyze empirical data, and formulate and evaluate data-based arguments. With proper modifications, the teacher can illuminate a variety of fundamental concepts in probability and statistics.

Engaging Different Intended Audiences

We have successfully used the *Schoolopoly* task with middle school students and argue that it is quite appropriate for use in high school or even college-level courses. Moreover, we have used the task with preservice and inservice middle and high school teachers as a means of exploring powerful connections between data and chance. Typically, traditional curricular materials compartmentalize the study of probability and statistics, placing the topics in separate sections, chapters, or units. *Schoolopoly*, however, requires students to synthesize their understanding of numerous topics to make decisions based on data.

Our experiences lead us to believe that most current and future mathematics teachers have had little, if any, experience with tasks such as this. In fact, the role of sample size in making inferences was typically not evident among teachers prior to engaging in this task. Consequently, this task serves to enhance teachers' content knowledge by connecting probability and statistics, and places them in a more favorable position to implement reform-oriented activities such as *Schoolopoly*.

Summary and Conclusion

In this chapter, we demonstrate that middle school students can engage in statistical inference using empirical data, recognize the importance of using larger samples in drawing inferences, and use data displays to make powerful connections between data and chance. Tasks like *Schoolopoly*, coupled with the social interaction among students and between students and teachers, provide for a potential meaning-making environment for the simultaneous development of probabilistic and statistical reasoning. As stated in *Principles and Standards for School Mathematics*, technology can "afford students access to relatively large samples that can be generated quickly and modified easily" (NCTM 2000, p. 254). Whether coupled with the use of available technology or utilizing hands-on materials such as number cubes, problem tasks such as the one described herein offer students opportunities to grapple with numerous issues central to the study and understanding of probability and statistics. In doing so, students can learn the value of formulating data-based arguments and recognize the importance of larger samples in drawing inferences.

References

Aspinwall, Les and Tarr, James E. “Middle School Students' Understanding of the Role Sample Size Plays in Experimental Probability.” *Journal of Mathematical Behavior* 20 (November 2001): 1-17.

Metz, Kathleen E. “Why Sampling Works or Why it Can't: Ideas of Young Children Engaged in Research of Their Own Design.” In *Proceedings of the Twenty-First Annual Meeting of the North American Chapter of the International Group for the Psychology of Education*, edited by Fernando Hitt & Manuel Santos, pp. 492-498. Columbus, Oh.: ERIC Clearinghouse of Science, Mathematics, and Environmental Education.

National Council of Teachers of Mathematics. *Principles and Standards for School Mathematics.* Reston, Va.: National Council of Teachers of Mathematics, 2000.

Stohl, Hollylynne. *Probability Explorer*. Software application distributed at http://www.probexplorer.com, 1999-2004.

Stohl, Hollylynne and Tarr, James E. “Developing Notions of Inference Using Probability Simulation Tools.” *Journal of Mathematical Behavior* 21 (November 2002): 1-19.

Watson, Jane M. and Moritz, John B. “Fairness of Dice: A Longitudinal Study of Students' Beliefs and Strategies for Making Judgments.” *Journal for Research in Mathematics Education* 34 (July 2003): 270-304.