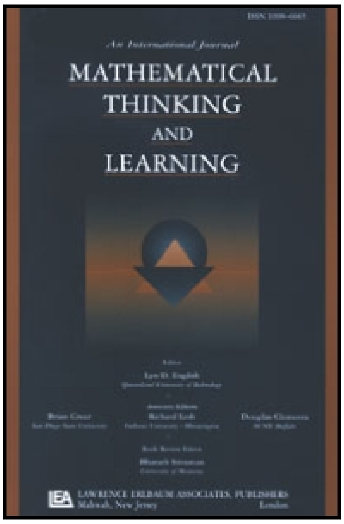

Mathematical Thinking and Learning
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653685

Conceptual Challenges in Coordinating Theoretical and Data-centered Estimates of Probability

Cliff Konold[a]; Sandra Madden[a]; Alexander Pollatsek[a]; Maxine Pfannkuch[b]; Chris Wild[b]; Ilze Ziedins[b]; William Finzer[c]; Nicholas J. Horton[d]; Sibel Kazak[e]
[a] University of Massachusetts Amherst, [b] University of Auckland, [c] KCP Technologies, [d] Smith College, [e] Pamukkale Üniversitesi,

Online publication date: 22 January 2011

To cite this Article Konold, Cliff , Madden, Sandra , Pollatsek, Alexander , Pfannkuch, Maxine , Wild, Chris , Ziedins, Ilze , Finzer, William , Horton, Nicholas J. and Kazak, Sibel(2011) 'Conceptual Challenges in Coordinating Theoretical and Data-centered Estimates of Probability', Mathematical Thinking and Learning, 13: 1, 68 — 86

To link to this Article: DOI: 10.1080/10986065.2011.538299

URL: http://dx.doi.org/10.1080/10986065.2011.538299

Mathematical Thinking and Learning, 13: 68–86, 2011

ISSN: 1098-6065 print / 1532-7833 online
DOI: 10.1080/10986065.2011.538299

Routledge
Taylor & Francis Group

Conceptual Challenges in Coordinating Theoretical and Data-centered Estimates of Probability

Cliff Konold, Sandra Madden, and Alexander Pollatsek
University of Massachusetts Amherst

Maxine Pfannkuch, Chris Wild, and Ilze Ziedins
University of Auckland

William Finzer
KCP Technologies

Nicholas J. Horton
Smith College

Sibel Kazak
Pamukkale Üniversitesi

A core component of informal statistical inference is the recognition that judgments based on sample data are inherently uncertain. This implies that instruction aimed at developing informal inference needs to foster basic probabilistic reasoning. In this article, we analyze and critique the now-common practice of introducing students to both “theoretical” and “experimental” probability, typically with the hope that students will come to see the latter as converging on the former as the number of observations grows. On the surface of it, this approach would seem to fit well with objectives in teaching informal inference. However, our in-depth analysis of one eighth-grader’s reasoning about experimental and theoretical probabilities points to various pitfalls in this approach. We offer tentative recommendations about how some of these issues might be addressed.

Erin[1] had been sitting with her hand raised at the back of the room for about a minute when we called on her. We had been discussing the various possibilities involved in drawing twice with

[1]Erin is a pseudonym.

We thank the students and staff at Lynch Middle School, Holyoke, Massachusetts, and Amy Robinson for the bone drawings. This research was supported in part with a grant from the National Science Foundation, ESI 0454754. The views we express here do not necessarily represent those of the Foundation.

Correspondence should be sent to Cliff Konold, Scientific Reasoning Research Institute, 424 LGRT, University of Massachusetts Amherst, Amherst, MA 01003-2545. E-mail: konold@srri.umass.edu

replacement from a bag that contained two plastic chips marked in different ways, and had been exploring what we meant by the claim that one of the events had a probability of 1/2.

Erin: Ok. Well, you know how you see it as half? Well, it doesn't really have to be half, because one's theoretical probability and the other one is experimental probability. And you're just looking at it from a theoretical probability. So it won't exactly be half, because if we actually *try* it and then do the data, and actually figure out the percents that way, it will come out to two different things. So before you could really assume that it's half, you need to do the data and see if it will actually come out half.

We were, of course, elated to hear this level of analysis from an eighth-grader. And as too often happens when a student says something intelligent in one of our classes, this was not something she had learned from us. This was the tenth day of an after-school program on probability and data, and we had not yet introduced this distinction.

I[2]: Where did you hear the names theoretical and empirical [experimental]?
Erin: Math class.
I: Math class? So you've had this discussion about this topic in math class?
Erin: No. Just what they meant. We didn't really go deep into it. We just, like, discussed it like that [snapping fingers to indicate a brief period of time].

Over the next few class sessions, Erin made similar statements that led us to believe that she had a fairly good sense of the relationship between probabilities computed from theory and those estimated from conducting actual trials. In addition, she seemed to be a committed empiricist in that she expressed reservations about theoretical probabilities, preferring information obtained from data. To allow us to further probe her thinking, she agreed to stay an hour longer one day, and for that occasion, we put together a series of problems and interviewed her as she answered them. Her responses to that interview comprise the focus of this article. At many points she surprised us, and our analysis of her interview has caused us to rethink some of our ideas about how we should introduce young students to probability.

One of those ideas relates to a practice, now common in the United States and other countries, whereby young students are introduced to both classical and frequentist interpretations of probability. In curricular materials, these two interpretations are typically referred to as "theoretical" and "experimental" probability, respectively, and these are the terms that Erin had learned in her mathematics class the previous year.

THEORETICAL AND EXPERIMENTAL PROBABILITIES IN THE CURRICULUM

In their review of the research on probability, Jones, Langrall, and Mooney (2007, p. 928) remarked that

> Given the importance of experimental (empirical) probability in national curriculum programs . . . it is surprising that so little research has been undertaken on students' conceptions of experimental probability, and "the bi-directional relationship between empirical and theoretical probability and the role of sample size in that relationship" (Stohl & Tarr, 2002, p. 324).[3]

[2]In the excerpts that follow, *I* indicates the Instructor/Interviewer.
[3]In Jones et al. (2007), this page number was mistakenly given as 314.

There are now a few studies that do explore the question of this “bi-directional” relationship. Several of them are reviewed in the recent article by Ireland and Watson (2009). Consistent with prior studies, which typically have employed an instructional intervention, these researchers reported that most of their students, aged 10–12, “were able to articulate, at some stage during the research, a basic level of understanding of the relationship between experimental and theoretical probability, mentioning a leveling out of results or ‘proving’ of one another” (p. 358).

The Law of Large Numbers proved the most difficult construct for Ireland and Watson’s (2009) students to grasp. The researchers attributed this to be due in part to “the lack of understanding of the implications of the underpinning theoretical concept of fairness” whereby many students continued to believe that fairness implied obtaining the expected distribution of results in the “short-term” (p. 359). This is a notion that Tversky and Kahneman (1971) observed even among statistically educated adults and dubbed the “belief in the law of small numbers” (p. 109).

Our interview with Erin, which we undertook to explore her understanding of the relationship between experimental and theoretical probabilities, in fact led us to question what we would even *mean* by these two conceptions of probability being related. Historically, the various schools of thought on probability (e.g., classical, frequentist, subjectivist) developed as reactions to one another, each attempting to put probability on a more sound foundation by addressing perceived flaws in the way previous theorists had defined probability (Mackie, 1973). It would no doubt perplex von Mises (1957), who developed the frequentist theory of probability, that we are now teaching both classical and frequentist-based ideas to students in a way that suggests that these two approaches to probability fit together hand-in-glove, because he saw himself as driving the nail in the coffin of classical theory. For him, probability is solely about repeatable phenomena, and in those contexts it is the limiting ratio of number of successes to number of trials as the number of trials goes to infinity. The question of whether the simple outcomes of a particular chance experiment are equally likely is “as irrelevant for our theory as is the moral integrity of a patient when a physician is diagnosing his illness” (p. 13).

While it was not the first to offer activities that introduced students to probability by drawing on both classical and frequentist interpretations, the *Exploring Probability* unit of the *Quantitative Literacy Series* was perhaps the most influential in promoting this approach (Newman, Obremski, & Scheaffer, 1987). These authors referred to the former as “theoretical probability” and the latter as “estimated probability.” The ideas and activities included in that unit are clearly discernible in many of the curriculum materials that followed (e.g., Singer, Konold, & Rubin, 1996 ; Bright, Frierson, Tarr, & Thomas, 2003; Lappan, Fey, Fitzgerald, Friel, & Phillips, 1998). At some point, these frequency-based estimates of probability began being referred to as “experimental” or “empirical” probabilities and defined in a way that obscured the fact that observed ratios were *estimates* of the probabilities, rather than probabilities themselves.[4]

There are several plausible pedagogical reasons for taking this two-pronged approach to introducing probability. Chief among them is the goal that students come to expect that the relative frequency of actual trials of some chance phenomenon will converge to the theoretical probability as the number of trials grows large. To facilitate this goal, activities often have students compute theoretical probabilities of situations like dice rolling and coin flipping and then compare these

[4]For one of many examples, see Billstein and Williamson’s (1998) Book 2 of *Middle grades MATH Thematics*, pp. 28 and 33.

probabilities to the relative frequencies they observe when they actually conduct (or simulate) trials.

Ideally, such activities also communicate the dangers of drawing conclusions from small samples, whose results will often deviate wildly from the expected outcomes. However, we are about to ignore that principle in this article and draw conclusions about teaching probability based on our interview with Erin—an *N* of 1. We therefore need to admit from the start that we are not under the illusion that if we repeated this interview with other students we would replicate what we report here.

BACKGROUND ON ERIN

Erin was in many ways a standout—precocious, deeply curious, playful, and quite conscious, even sometimes boastful, of her intellectual prowess. On the 2006 Massachusetts Comprehensive Assessment System (MCAS), Erin's scores on the English and Mathematics components put her about midway between the "proficient" and "advanced" levels as specified by the state. Within her school (which is a low performer within the state) she scored at the 76th percentile in English and 94th percentile in Math. During our after-school program, she preferred to sit in the rear of the room apart from the rest of the students and was more interested in engaging us than in engaging her peers. Thus Erin in many ways is not a "representative" case.

In the end, however, this single interview was compelling enough to raise serious questions for us about how we introduce students to probability. What made it so was the fact that we came to see Erin's particular understanding of probability as a logical consequence of the way she had been introduced to probability. In our view, she had attended carefully to the definitions of experimental and theoretical probability that she had been offered. She attempted to integrate these ideas with other beliefs she held to make better sense of probability. Thus, her understandings came to represent to us how students might and should reason if they attended carefully to what much of the current curriculum attempts to teach them about probability. The fact that we know that many students will not pay such close attention to what instruction offers them is not something, pedagogically, on which we should be counting. It is this view of Erin as an ideal learner rather than a representative one that, for us, makes this single interview a worthwhile source of reflection.

There are two main characteristics of Erin's formal introduction to theoretical and experimental probabilities that we conclude posed conceptual difficulties for her. The first is the definition she learned for experimental probabilities. The sixth grade unit in the Connected Mathematics Project (CMP) materials adopted at her school defines experimental probability as the ratio of the number of favorable trials to the total number of trials.[5] Defined in this way, experimental probability is bound to change with each sample. As we indicated, many materials we have looked at in the United States now define experimental probability in this same way, omitting the critical piece of information that this ratio is an *estimate of something*—of the actual, but ultimately unknowable, probability of the outcome of interest. Hereafter, we will refer to this latter probability as the "true probability."

[5]See p. 1b of the Teacher's Guide of the grade 6 unit "How Likely Is It?" (Lappan, Fey, Fitzgerald, Friel, & Phillips, 1998).

The second characteristic of Erin's introduction to probability is found in many of the published activities we have examined where the materials seem to make a tacit assumption that the theoretical probability (i.e., the probability one can compute assuming an idealized model—usually that certain elementary events are equally likely) is the true probability, and thus the value on which the experimental probability will converge. But, as we will see, Erin does not make this assumption, and we think it reasonable that she does not. As a result, there is no anchor for the experimental results she observes—nothing for the observed experimental probability to converge on. While there are other aspects of Erin's reasoning that pose challenges for her development of a richer view of probability, these two shortcomings in her formal instruction—an incoherent definition of experimental probability and no explicit development of the idea of an underlying but unknowable true probability—appear to serve as substantial roadblocks to her progress.

As mentioned, Erin was participating in an after-school club in which Konold and Kazak served as instructors. The club met for an hour once a week during the 2006–2007 school year at a middle school in Massachusetts, USA. Twelve students in grades 6–8 participated, although attendance from week to week was quite variable. The general focus of the after-school club was data analysis and probability. We were testing activities we created that made use of a new version of *TinkerPlots* we were developing. This new version (version 2), adds the ability to design and run probability simulations to the data-analysis capabilities of *TinkerPlots* 1.0 (Konold & Miller, 2005).

At the point of the interview we report next, Erin had participated in 13 club sessions, 8 of which had focused on probability. For the most part, these sessions had involved students in learning to build computer simulation models of chance situations and examining data from them. In Konold and Kazak (2008), we described the instructional approach that eventually evolved as a result of this and subsequent field tests. The primary emphasis in the "chance" activities was exploring features of chance distributions rather than computing probabilities of particular events. The session described in the introduction of this article, which occurred on the tenth meeting of the club, was the first time we had asked students to compute and interpret a probability. We had not intended during this session to introduce students to the terms "experimental" and "theoretical"; as we indicated, these terms were introduced into the discussion by Erin.

THE INTERVIEW

For much of the interview, Erin used the beta version of *TinkerPlots* 2.0. By the time of the interview, Erin had experienced several hours of use of this software. This included building and running simulations and analyzing the results. We developed four problems for the hour-long interview. We will describe her responses to the first three of these in the next section.[6] The interview was semi-structured so that while many of our follow-up questions were planned, we also asked spontaneous questions to further probe her responses.

[6]We do not discuss Erin's responses to the fourth and last task because our analysis of it suggested that she was confused about the graph we had given her to reason from, one that attempted to show the percentage of 2s from a die model settling down as the number of trials grew.

Spinner Problem

In this problem, we asked Erin to determine the probability of getting green from the spinner shown in Figure 1. The spinner was divided such that 30% was green and 70% red. However, details of the spinner were hidden when we first showed it to her; we told her only that it was part green and part red. She had previous experience with devices of this type in *TinkerPlots*, where the task was to infer the contents of a masked sampling device by “running it” and analyzing the resulting data. But in these prior cases, she was trying to guess the shape and approximate centers of hidden frequency distributions. This was the first time that we had asked her to infer an unknown probability from such a masked device. On this occasion, we first asked if she had an initial idea of how the spinner might be divided. Then we had her conduct trials by running the spinner with an interest in seeing what inferences she would draw from the graphs of the outcomes of these trials (see graph in Figure 1). Finally, we revealed the contents of the spinner and then had her draw more samples. Throughout these stages, we asked her what she thought the probability of green was and probed her about how she was arriving at her conclusion.

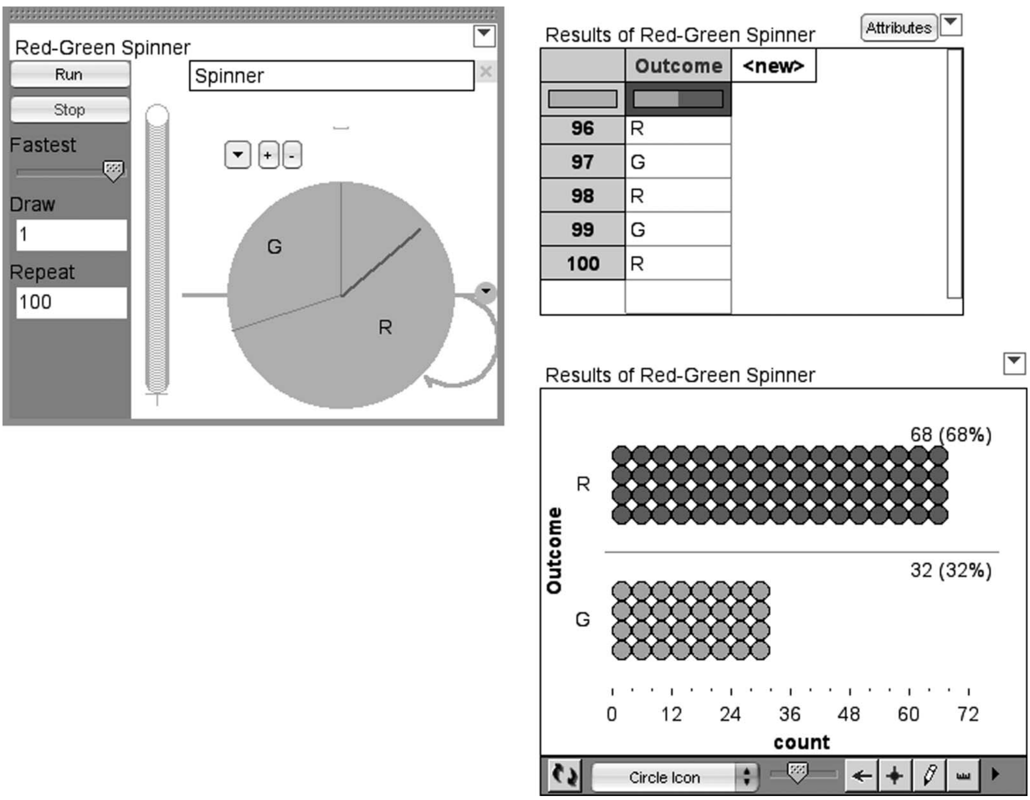

FIGURE 1 The contents of the spinner at the upper left were originally obscured. Erin could draw samples of any size by changing the Repeat number and hitting Run. The graph in the lower right was set up to show the number (and percentage) of the two outcomes.

This problem required Erin to explore a situation experimentally before she saw the structure of the outcomes from which she could derive the theoretical probability. Our expectation based on what she had said in prior class discussions was that while she would use data to make guesses about the probability of green, once we revealed the hidden spinner she would then say that the probability of green was 30%. That is, we thought her distrust of theoretical probabilities was based on her questioning whether physical objects such as dice were truly symmetrical or whether we could conduct trials on them using truly random methods. This presupposes, of course, that it is not unreasonable to assume in this instance that the theoretical probability (the percent of the spinner that is green) and the true probability are the same. Although there is no principled reason to believe in this instance that the true probability of getting a green is actually 30%, our experience is that students generally accept it as such from a computer more readily than from a physical device. Erin had expressed her distrust during classroom discussions that real coin flipping and dice rolling were perfectly "fair," but did not offer any such objections about the computer simulation. Indeed, later in the interview while introducing the Die Problem, we posed this question directly to her.

I: So then which would be fairer to use? This die model [built in *TinkerPlots*], or a real die, if you wanted to be fair?
Erin: That model [pointing to computer spinner model].
I: Because?
Erin: How will you. . . . People won't cheat and actually say, you know, "I'm going to try to get more 2s, or more 5s."
I: Alright. Suppose we didn't allow people to do the kind of cheating you talked about. We put it in a cup and we made them shake it real well and then roll it. Then would this data . . .
Erin: No, because also it's depending on, like say if you. . . . Like, a number will be more likely. Say if this [her open palm] was a cup and you put it right there. It's also depending on the number it starts at on top.
I: Ok.
Erin: I think [tone suggesting some uncertainty].
I: So you'd trust this [points at computer model] to be fairer, more than you would a real die, no matter how you rolled the real die?
Erin: Mm huh.

For the Spinner Problem, we initially asked Erin: "What do you think is the probability of green?" Her initial answers focused primarily on why we were unlikely to have divided the spinner in half ("that would be too obvious"). However, she also explained that no matter what the actual percentage of green was on the spinner, "you can't be sure on any set of data about the results you will get."

Erin: Even an outcome that is 25% chance can turn out more than the outcome that is 75%.

Next, she drew samples of 100, repeating this 5 times (the number of greens were 24, 34, 27, 25, and 34). At the end of these trials, her best guess was "one third, or 33%."[7] We then revealed the spinner and asked:

[7] One third is obviously (to us) on the high side of her results. She may have chosen that estimate under the assumption that we surely would have divided the spinner using some common unit fraction: "school" usually works like that.

I: And so what's the probability of green?
Erin: 30%. [When we revealed the spinner, a feature was turned on that displayed this percent directly on the spinner.]
I: And you're getting that 30% from what?
Erin: From the graph [points to spinner].

The fact that she said 30% (rather than "about 30%") could be an indication that she had the concept of a single theoretical probability that the sample proportion was estimating. However, most of her subsequent responses, including the following, make this interpretation unlikely.[8]

I: And why do you take the probability from the spinner rather than from the graph down here [points to graph of results of 100 trials].
Erin: Because the circle is divided up into two parts, so it will tell you more of what it is, because this [points to graph] isn't as—experimental and theoretical probability. This [points to spinner] is a theoretical probability and the graph is an experimental probability.
I: I didn't ask you "What's the theoretical" or "What's the experimental probability." I just said, "What's THE probability." And you're giving the answer as . . .
Erin: You could get it two different ways.
I: Uh huh. So, and which way do you believe in most in this case?
Erin: More of the experimental, because if you do it by experimental you're going to get more accurate results, because there's more than, like if you go by theoretical, you won't make it as much, and . . .
I: You wouldn't make "it" as much?
Erin: Yeah, because say like 30% and 70%, you won't make green 30% of the time, so it's not really accurate to when you actually try it out.
I: So this 30% [the percent showing on the spinner] isn't going to be as accurate as what you get when running it down here [the results of a trial showing in the graph]?
Erin: It will be CLOSE, mostly around the same distance. Like it might be—like one time it might be 29% of the time, but another one might be 33%.

Our analysis of this and several similar statements Erin made later during the interview suggests that a major reason that she "believes in" experimental probability more than theoretical probability is because it reflects the variability she observes when she repeatedly conducts trials. Thus, she seems to clearly have the two concepts:

1. Exactly 30% of the spinner is green and the computer is unbiased in its selection, and
2. The observed sample proportion is unlikely to deviate a large amount from 30%.

However, it is not clear that she has the idea that the proportion she observes in a particular sample is an estimate of this 30% value.

Assuming that there is a coherent view underlying her responses, the obvious question is: What is Erin's conception of probability? The tentative view we offer here is that she regards the 30% value visible in the spinner as a guide to predict what will happen with actual trials rather than as the underlying, true probability. One possible reason for her not wanting to regard the 30% as the true probability value is that if she uses this value as her prediction of what will occur when she samples, she will rarely be (exactly) right. Indeed, with certain sample sizes (e.g., 25 in

[8]The words that are capitalized below were stressed in the interview.

this example), the expected number (7.5) is impossible. A second reason may be that a specific value does not capture the variability that she knows occurs over repeated trials. Thus, our sense is that she does not view the collection of experimental sample proportions as values that will tend to cluster around a true value. Rather, she perceives them as the fundamental reality—as "what's really happening."

Our interpretation is that Erin is uneasy about using a single number to describe a probability because any specific proportion is unlikely to occur in a sample. Instead, she describes experimental probabilities as "more accurate" because they better capture what will happen (i.e., they are not the same from sample to sample). At various times she describes these probabilities as a range of possible values that will stay "mostly around the same distance" from one another or, as noted next, as single values that change from trial to trial.

I: So does the probability of green [with the red/green spinner] change every time we run this?
Erin: Almost, yes. But there are still times that it's the same.
I: What time would it be the same?
Erin: Like if you run it twice and it's both 32% the same.

To summarize, from her answers to the Spinner Problem it appears that Erin views the theoretical and experimental probabilities to be related, but the relation is vague. In some sense, she seems clear that the observed proportion on each trial run will not be too far away from the theoretical probability and thus that the theoretical probability conveys information about the situation. But we see no evidence that she views the theoretical value as the true probability that the experimental results cluster around and provide estimates of. The fact that she seems to have the notion of "experimental probability" as being a collection of values may be getting in her way of thinking of a single sample proportion (which is what one generally collects in real life) as an estimate of an unknown but fixed population value. Her responses during the remainder of the interview both reinforce this tentative conclusion and reveal other interesting aspects of her views of theoretical and experimental probability and the relation between them.

Die Problem

For the second problem, Erin explored a spinner model of a fair, six-sided die built in *TinkerPlots*. After first presenting her a real die and then showing her the model, we asked her:

I: So what's the probability of getting a 2 with this model in *TinkerPlots*?
Erin: Same thing as the [real] dice. Around 16.67%.
I: And why do you say AROUND?
Erin: Because, for one, like, in real life, the probability of things, you can't really go out to a decimal, and since in this probability you have a decimal. And also that if you actually do it experimental, it will be around 16%, not exactly 16% every time.

This again suggests that her discomfort with the theoretical probability was that you rarely get this value (here 16.67%). In this case, the fact that one cannot get exactly 1/6 2s in 100 rolls seems to loom large in her mind. In addition, she again focuses on the fact that the actual value of the proportion of 2s varies from trial to trial.

After conducting numerous trials with the spinner model of the die, we further probed her thinking about the probability. The last trial of 500 rolls had produced 18.6% 2s. We asked:

I: Would you agree that the experimental probability is 18.6%?
Erin: No, because how it changes every time. See, now it's 18.6%, but then when I do it another time [hits run], it is 16.4. So it changes every time.
I: Ok. And what's changing?
Erin: The probability of green, I mean 2.

Here it appears that her view of experimental probability is close to the definition that she had learned from her math class: the ratio of the number of successful occurrences to the number of trials. This definition specifies the ratio as *the* experimental probability rather than an estimate of the true probability. Thus, when that ratio changes with a new sample, the experimental probability also changes. But note that she balked when we asked her whether a particular value was *the* experimental probability. This suggests that she does distinguish between the percentage observed from a particular trial and some underlying probability. She does not see the 18.6% as *the* experimental probability even though it corresponds to the definition she has learned. So instead, she uses the observed frequency to make a claim about the neighborhood in which the experimental probability lies.

To determine whether she had some underlying notion of a true probability, we probed:

I: And so if you really wanted to know what the probability was—that doesn't change—what would you say it was and how would you find out?

As suggested in her response, she initially proposed averaging several experimental probabilities, but this did not lead her to an underlying, fixed probability.[9] She then offered that a theoretical probability provides a single value, but again expressed a lack of confidence in this approach.

Erin: I'd find out by, like, taking all the experimental probabilities, adding them up, and then div. . . . Like say if I had 5 probabilities, and then dividing it by 5 and by finding the average, and then I'd say that's about the probability.
I: So you think that's the way to get the real probability?
Erin: No, it's just going to find a number it's going to be around.
I: Is there any way to find the exact probability?
Erin: If you do it theoretically, yes.
I: But do you trust that theoretical probability?
Erin: Sometimes, or sometimes not, depending on if I'm too lazy to do the work, or too lazy to think.

[9]Ireland and Watson (2009) reported similar findings with a student they interviewed who thought it might help to compare results across several replications and yet apparently did not see this as the same as drawing one large sample (p. 355).

I: So how about in this case? You told us the theoretical probability is 16 point—or is exactly one sixth. So we can use that fraction so we don't have that funny decimal. So is the probability of rolling a 2 one sixth, or is it something else?
Erin: It's just about—it will stay around one sixth because of how there's six sides and you only want one side of it [looking at the actual die].
I: But you're not comfortable saying the probability IS one sixth, it doesn't sound like?
Erin: I'll say it's ABOUT one sixth.

Bone Problem

The third problem involved determining the side most likely to land upright for an irregularly shaped object—a deer's ankle bone (see Figure 2). In this situation, one cannot determine theoretical probabilities based on object symmetry. This, like the Spinner Problem when the spinner was hidden, is a paradigm case where the true probabilities are fixed but unknown values. Collecting data offers a way to estimate the probability.

After showing Erin the various labeled surfaces of the bone, we asked her:

I: If you were to roll this, which side do you think would be most likely to land upright?
Erin: Can I see it [the bone]?
I: Yeah [handing her the bone].

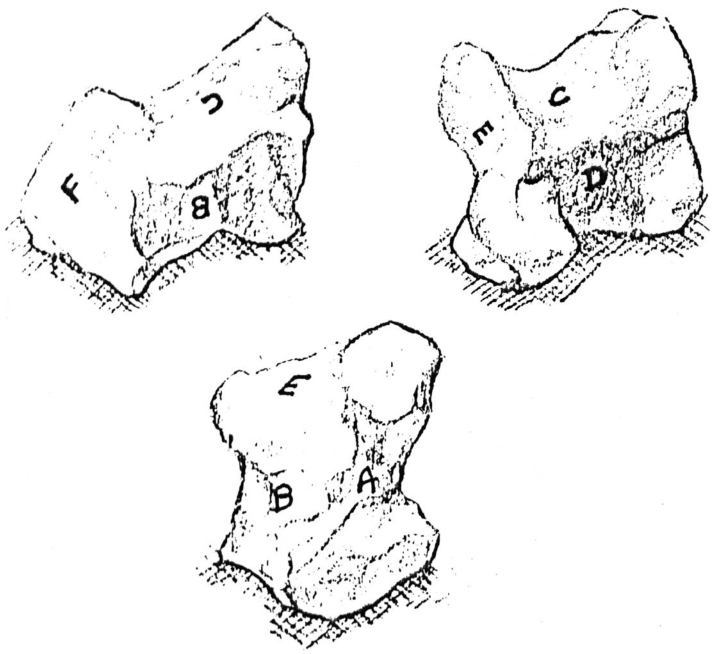

FIGURE 2 Three perspectives of the deer bone we gave Erin, showing the bone's six labeled (A–F) surfaces.

Erin, like many adults given this problem (Konold, 1989), made some reasonable inferences based on her inspection of the bone. Sides E and F are at the ends of the longer axis, and thus most people quickly judge them as being relatively unlikely. Furthermore, the remaining four sides are different enough that most people choose sides D or B as being most likely (which, indeed, they are). Inspecting the bone, Erin commented:

Erin: F is very unlikely because of how it cannot really stand up. . . . Probably either a D or a B, more likely a D, probably.

She was eager to roll the bone, and had actually rolled it once the moment we handed it to her. Our expectation was that after rolling the bone that she would use this as an occasion to elaborate her reasons for preferring the experimental approach to probability. In this context, it is the only way one could proceed to get reasonably accurate estimates of the probabilities of the various sides and therefore to determine which side was most likely. After eliciting her judgments based on inspection, we suggested that she roll it again.

I: Why don't you go ahead and roll it and see what happens.
Erin: And I'll be wrong. [Rolls a B.] B. It's a B or a D. [Rolls another B.] B.
I: So, wait a minute [as Erin rolls again]. So what do you think so far about your prediction?
Erin: I was right.
I: What do you mean?
Erin: It more lands on B or D. Because how when you roll it, B and D are the flat surfaces, so if it bounces, it will spin and then it will like find a flat surface that's easy to balance on. So B or D.

This interchange seems to suggest that she was reasonably satisfied with giving qualitative answers of the form that particular events were quite unlikely while others were more likely. Although gathering such qualitative assessments in such a situation is not unreasonable, we wanted to determine whether she thought of these probabilities in a more quantitative way, and if so, how to determine them.

I: And if you wanted to know which was more likely, B or D, how would you find out?
Erin: Looking at it [picks up bone and inspects]. Probably . . . B is more likely, because when you drop this and it lands on B, see how this has, kind of right here, how it has like a big spot? Like when you drop it, it will bounce off from that and then it will flip over in the air and land on B.
I: And suppose you wanted to find out and say what the probability of B was. Suppose you wanted to find that out. What would you do?
Erin: Look at it. Find the probability out of each one and then like kind of come to an agreement.
I: What do you mean "find the probability out of each one"?
Erin: Because, like E. It's not that likely. So it would be about like one eighth [rotates bone to reveal another side]. This would be about like one. . . . No, wait [rotates back to

previous surface]. This would be about like one fifth. [Rotates.] One seventh. [Rotates to new surface.] And then like that. I'll find something like that.

I: So if we gave that [bone] to you to take home and come back tomorrow and give us a good, accurate estimate of what the probability of B was, what would you spend the day—what would you do at home?

Erin: Ask my brother. Say, "Ryan, do this," and he'd do it.

Apparently, it did not occur to her to roll the bone a large number of times and use those trials to compute experimental probabilities. We had initially predicted that with the Bone Problem, this approach would seem to her the only sensible way to proceed. So we repeated our request after stressing that we would want a very accurate assessment of the probability of B and that we would give her one day to come up with it.

I: So start your investigation. What would you do?

Erin: Ask my math teacher, because he's very good with math.

I: And what kind of math would he use to figure that out?

Erin: Probability math [said jokingly].

I: So that's what you'd do—you'd go ask the experts, huh?

Erin: Uh huh.

We next showed her results we got from rolling the bone 1,000 times, and she looked incredulous that we would have actually rolled it that many times.[10] However, she did use those data to support her expectation based on physical inspection that B and D would be the most likely sides to land upright. When we asked her to, she went on to use the data from 1,000 rolls to come up with a figure of "approximately 28%" for the probability of B. But when we pressed her again for how she might get an even more accurate estimate, she was doubtful that more data would provide that information.

I: And so again, suppose I gave you a day to get a very, very accurate estimate [of the probability of B]. What would you do?

Erin: Ask you. You know.

I: Well, what about rolling the thing a whole bunch of times and keeping track?

Erin: My hand will get too tired. And also because of how it's more experimental it wouldn't really work [as she is rolling the bone] because no one has time to do it 1,000 times.

I: I did!

Erin: Not me. I have a very busy schedule.

I: But you could pay someone to do it. We're going to pay you 1,000 dollars. You could go ask your brother and pay him some percentage. . . . Would that help? If I rolled that like a whole bunch more times and kept track, would that help me get the probability?

Erin: More—probably if you did it—probably, yeah. If you rolled it more than 1,000, because it might be like a little more accurate. Or if you did a lot of different sets and then got their average, and then it might be a little more accurate.

[10] These frequencies were A = 50, B = 279, C = 244, D = 375, E = 52, and F = 0.

She had expressed this idea of averaging averages previously in the classroom as well, and at that time we wondered if she had a basic understanding of the Law of Large Numbers. It seems clear from her statement here, however, that she views the additional accuracy from conducting more trials as resulting primarily from the act of averaging these averages, rather than from the increased sample size. Thus we would predict that if we asked her whether she would get better information from looking at the percent of Bs in 1,000 rolls or from looking at the averages of the percentages obtained in 10 repetitions of 100 rolls, that she would prefer the later.

The tone of Erin's previous response suggested that she was skeptical about the benefits of collecting more data, and we probed her as to why.

Erin: Because I don't know how to do it. [Picks up bone.] Because how like each of them [each side] has a different probability of each one.
I: You gave us a probability based on this [pointing to numbers in table of 1,000 rolls], and I assumed that you'd think that was an experimental probability.
Erin: [Nods slowly, but looks unsure.]
I: Or is it something different?
Erin: Experimental.
I: So is there a way to get a theoretical probability for this thing?
Erin: [Shakes head no.] Not that I know of. But you're probably about going to tell me.
I: No. I don't know how to do it.

Erin certainly had the idea that rolling the bone would give her useful information. She spontaneously rolled it at the first opportunity, and once we asked her to roll it, we had to intervene to stop her from rolling it several times before we could probe her thinking. Furthermore, she used results from her rolls and our 1,000 trials to support and somewhat modify her prior expectations. But when we asked how she would go about generating a very accurate estimate of the probability of side B, she had no idea about how to proceed. She assumed that experts would have some way of getting these probabilities; however she apparently did not believe that they would do this by rolling the bone a larger number of times. We conclude that what she was lacking was the idea that the bone has a true probability of landing B and that results from trials would allow her to zero in on what that true probability is. Indeed, to see how accurate probabilities could result from conducting trials requires understanding the Law of Large Numbers.

Interview Summary

What was missing in Erin's concept of probability was the notion of a true probability—the idea that associated with a chance outcome is a single value of probability that we can model or estimate but can typically never know precisely. This is an idea that commonly remains unstated in the theoretical-experimental approach to instruction. We think that curriculum designers assume that the student will accept the theoretical probability as the true probability, but this assumption is unwarranted because there are really no good grounds for believing that any real coin is exactly equally likely to land head or tails. With virtual objects like the red/green spinner, it may be tempting to believe that the theoretical and true probabilities are the same, but even this rests on beliefs about the computer's random number generator.

Erin did not appear to view the theoretical probabilities she calculated as true probabilities. Indeed, she distrusted theoretical values because they do not, in general, predict what really happens. This perception may be supported by the term "theoretical," which we think she interpreted roughly as "this tells you what *in theory* ought to happen, but experimental probability tells you what *really* happens."

Moreover, while she preferred the information she obtained from conducting trials, she did not yet have in place the ideas: (a) that there is some true probability (even if it differs from the theoretical value) associated with the phenomenon, and (b) that the more data she collected, the closer she could expect her estimate based on that data to be to that true value. For this reason, she could not see a way in the Bone Problem to get more accurate estimates of the probability of B from collecting more data. Rather, she suspected there was some other method that an expert would use to construct probability information from the bone (e.g., calculating percentages based on somehow knowing the precise area of each face).

Erin's preference for experimental probabilities over theoretical might be an adaptation of the "outcome approach" (Konold, 1989). In this approach, people see the objective of probability as making predictions about what will happen on a next trial—predictions that can be evaluated as being correct or incorrect after that trial. Erin was generally thinking not about single trials but about samples, and in this regard, was not reasoning according to the outcome approach. However, she still appeared to view the objective in probability as predicting what will happen. From this viewpoint, the theoretical value usually fails, because you almost never get that exact value appearing in a particular sample. On the other hand, the experimental probability, expressed as a collection of possible values, is more often "verified" in a particular experiment.

CURRICULAR IMPLICATIONS

Recently, members of the statistics education community have suggested that introductions to data analysis in the K-12 curriculum be oriented toward developing a set of "informal inference" skills. Contributors to this special journal issue are among the strongest proponents of this reorientation. Although the various proposals for teaching informal inference differ somewhat in their conception of what informal inference entails, all of them include the idea that students associate some probability-based notion of uncertainty with inferences they draw from data (e.g., Makar & Rubin, 2009). In this respect, these recommendations are a reversal of those made several years ago by Moore (1992) and others (e.g., Watkins, Burrill, Landwehr, & Schaeffer, 1992) that we limit study of probability in introductory statistics courses or even remove it entirely. In light of this rethinking, some researchers have begun exploring how we might integrate instruction in data analysis and probability to support one another (e.g., Konold & Kazak, 2008).

As we pointed out, prior research as well as many current curriculum materials that aim to connect experimental and theoretical probability have tended either not to mention or to gloss over a notion that we think is fundamental and directly supports the objective of students learning to make inferences *informally.* That notion is this: when we consider the probability of some event, we should be thinking of a specific, single value that we, in general, can never know. It is this value, which we have been referring to as the *true* probability, that we are, in fact, interested in knowing. This true probability is not ordinarily synonymous in real-life situations with a theoretical probability. Rather, a theoretical probability is derived from a model that we sometimes can apply to a real situation, a model which is, at best, a good approximation to the real situation. One way to determine how well the model fits is to collect data from the real situation and compare the estimate of true probability with the theoretical probability.

This set of relations among theoretical (or model-driven) probability, the true probability, and estimates of the true probability from frequency data, are the ideas and relations that, in our opinion, should be the focus of introductory instruction in probability. From this perspective, it is misleading and incorrect to claim that data from rolling a real die will converge on the theoretical probability, which is how many curricula currently portray it. Instead, it converges on the true probability of that particular die. It will only converge (also) on the theoretical probability if that happens to be a perfect model for that die.[11]

We suggest that the current practice of having students explore probabilities associated with situations such as die rolling and coin flipping from both the theoretical and experimental perspectives is not well suited to highlighting the idea of an unknown but true probability. This is primarily because these explorations tacitly assume that students will accept the theoretical analysis as synonymous with the true probability. Yet, many students express their doubts about this assumption in real situations. And well that they should. Data from over 26,000 rolls obtained by Labby (2009) using a custom-built mechanical rolling apparatus and image-analysis software suggest that dice (excluding those carefully produced for use in casinos) are in fact slightly biased in favor of 1s and 6s. But our argument here is not that a theoretical analysis is inappropriately applied to dice, coins, and spinners. Our point is that by tacitly assuming the theoretical model, current curricular materials obscure the idea that what we really strive to know is the true probability. In many situations, the theoretical approach provides a model that may be reasonable; however, results from actual trials will converge on the true probability, whatever that is.

Furthermore, by encouraging students to view experimental results as converging on theoretical probabilities, we believe an important opportunity is lost to develop fundamental skills that are at the heart of informal statistical inference. When students think from the start that they know the true probability (i.e., that it is the theoretical probability), then the idea of uncertainty and confidence in one's inference—indeed, the very idea of an inference—is lost. When these students later move to investigations of naturalistic data, there are no clear theoretical or known values for sample statistics to converge on. This disconnect has bedeviled statistics education for years, leaving many university students to wonder what the probability chapter in their introductory statistics textbook had to do with anything that followed.

An alternative approach to probability instruction that we think holds promise is for students to *start* with explorations of situations such as the bone rolling or thumbtack tossing where there is no clear theoretical model.[12] In exploring these situations, students would see their objective as estimating a probability by collecting data and that the data gave them some information about the tendency of the object to land in particular orientations. These activities would serve as the basis for students later coming to understand that while the data we collect can inform our estimates of probability, we can never know exactly what that probability is for the same reason

[11]In practice, of course, it may be impractical to roll the die enough times to detect any difference between the true probability and the theoretical probability.

[12]The CMP unit *Data About Us* that Erin used does include more than one such activity, for example, flipping marshmallows and noting whether they land on an end or a side (see pp. 14–15 of the Teacher's Guide).

that we can never know the exact length of a table we measure, no matter how finely calibrated our ruler.

A limitation of explorations of physical objects is that we cannot ordinarily collect enough data in the classroom to observe the fact that our probability estimates from larger samples are less variable than those from smaller samples. We therefore suggest that, after these initial activities with physical objects, we move students to exploring computer objects such as the red/green spinner where students can collect and quickly analyze large amounts of data. In this environment it would be possible, for example, for students to collect enough data in a few minutes so that they could distinguish a two-sector spinner that was divided 71/29 from one that was divided 70/30. Activities such as this would be aimed at developing the notion that as a sample gets larger, the interval one can give of where the true probability lies can become narrower. Our objective would be for students to come to understand that while we can never know the true probability, we can narrow the interval where we expect it to be at a particular level of confidence. Our expectation would be that applying this same logic and reasoning would not seem such a leap when these students with this understanding moved on to data analysis and exploring random samples of populations. If they have this basic understanding, only the addition of a method of quantifying uncertainty would seem needed for them to understand the basics of formal statistical inference. However, we should add that what might seem a small step or logical application of an idea often proves to be a formidable conceptual challenge.[13]

Only after these ideas were established would we begin to involve the students in developing theoretical models from sample spaces. But here we would stress that these models for predicting true probability may or may not be very good. To make this point, we would include, along with examples where the model holds rather well, situations where students' initial expectations based on apparent symmetry are, with data, proved wrong.[14] Having them spin (rather than flip) coins is one such context, where the spinning takes advantage of the slight asymmetry in real coins and causes, with some coins at least, one side to land upright considerably more often than the other. A similar result is obtained by rolling wooden, hexagonal pencils the sides of which have been labeled 1 through 6. The results of 100 rolls are enough to convince most people that pencils do not behave like dice, as one or two sides of the pencil will occur much more frequently than the others. Careful analysis of the pencil reveals the reason—pencils are always at least slightly warped.[15]

CONCLUDING REFLECTIONS

In this article we have examined the conceptual challenges students face in coordinating theoretical and data-centered estimates of probability. Our reflections on one eighth-grader's

[13] Indeed, there is evidence from both the history of the development of statistical thinking and from classroom research that suggests that applying, for example, the idea of average as signal across different contexts is not a simple matter (Konold & Pollatsek, 2002; Lehrer & Schauble, 2007). Thus, taking ideas of informal inference developed in the context of probability and applying them to random sampling from populations of people (for example), will likely require carefully designed instruction.

[14] See the activity developed by Rider and Stohl Lee (2006).

[15] If you have collected data on ten pencils, you can blindly draw one from the collection, roll it a few times, and correctly identify which pencil it is. Each pencil has its own unique signal distribution.

reasoning have led us to delve into the nature of the relationship between theoretical and experimentally estimated probability. A point on which classical and frequentist theorists agree is that probability is fundamentally about the objective world (in contrast to subjectivist theorists who maintain that probability is a gauge of belief). This is the common ground on which the experimental and theoretical instructional approaches should build—that what we are interested in knowing is the true probability of e.g., the bone landing on side B or the die coming up 2.

The theoretical and experimental methods provide different ways of giving us information about what this true probability might be. The experimental approach is the more general, in that we can make headway with it in many situations where the theoretical approach is difficult or impossible to apply. Furthermore, by purposefully identifying experimental probability as an *estimate* of the true probability of the chance set up under investigation, we make strong links to making inferences informally.

REFERENCES

Billstein, R., & Williamson, J. (1998). *Middle grades MATH Thematics* (Book 2). Evanston, IL: McDougal Littell.

Bright, G. W., Frierson, D. Jr., Tarr, J. E., & Thomas, C. (2003). *Navigating through probability in grades 6–8*. Reston, VA: National Council of Teachers of Mathematics.

Ireland, S., & Watson, J. (2009). Building a connection between experimental and theoretical aspects of probability. *International Electronic Journal of Mathematics Education*, *4*(3), 339–370.

Jones, G. A., Langrall, C. W., & Mooney, E. S. (2007). Research in probability: Responding to classroom realities. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 909–955). Charlotte, NC: Information Age

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, *6*, 59–98.

Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, *2*(1), Article 1.

Konold, C., & Miller, C. D. (2005). *TinkerPlots: Dynamic data exploration*. Emeryville, CA: Key Curriculum.

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.

Labby, Z. (2009). Weldon's dice, automated. *Chance*, *22*(4), 6–13.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998). *Connected Mathematics*. Menlo Park, CA: Dale Seymour.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. Lovett & P. Shah (Eds.), *Carnegie symposium on cognition: Thinking with data* (pp. 149–176). New York, NY: Lawrence Erlbaum.

Mackie, J. L. (1973). *Truth, probability and paradox*. London, UK: Oxford University Press.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105.

Moore, D. S. (1992). Teaching statistics as a respectable subject. In F. S. Gordon & S. P. Gordon (Eds.), *Statistics for the twenty- first century* (MAA Notes, #26, pp. 14–25). Washington, DC: Mathematical Association of America.

Newman, C. M., Obremski, T. E., & Scheaffer, R. L. (1987). *Exploring probability*. Palo Alto, CA: Dale Seymour.

Rider, R., & Stohl Lee, H. (2006). Differences in students' use of computer simulation tools and reasoning about empirical data and theorectical distributions. In A. Rossman & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics: Working Cooperatively in Statistics Education, Salvador, Brazil* [CD-ROM]. Voorburg, The Netherlands: International Association for Statistical Education and the International Statistical institute.

Singer, M., Konold, C., & Rubin, A. (1996). Between never and always. In *Investigations in number, data and space*. Palo Alto, CA: Dale Seymour.

Stohl, H., & Tarr, J. E. (2002). Developing notions of inference using probability simulation tools. *Journal of Mathematical Behavior*, *21*(3), 319–337.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Review*, *76*, 105–110.

von Mises, R. (1957). *Probability, statistics and truth*. London, UK: Allen & Unwin.

Watkins, A., Burrill, G., Landwehr, J., & Schaeffer, R. (1992). Remedial statistics?: The implications for colleges of the changing secondary school curriculum. In F. S. Gordon & S. P. Gordon (Eds.), *Statistics for the twenty- first century* (MAA Notes, #26, pp. 45–55). Washington, DC: Mathematical Association of America.