

Running Head: MAKING CONNECTIONS BETWEEN EMPIRICAL AND
THEORETICAL PROBABILITY

Making Connections Between Empirical and Theoretical Probability:
Students' Generation and Analysis of Data in a Technological Environment

Hollylynne Stohl, North Carolina State University

Robin L. Rider, East Carolina University

James E. Tarr, University of Missouri

Submitted for review, September 2004.

DO NOT CITE WITHOUT PERMISSION

Abstract

In this study, we examine how sixth-grade students use empirical data to make inferences about unknown probabilities and make decisions about the fairness of dice. Students used a microworld environment to display and conduct simultaneous data collection and analysis. We examined the ways in which students' understanding and use of sample size, independence, fairness, and variability interacted with their use of external resources such as the task context, multiple representations of data, and social negotiations with a partner. Students who ran larger numbers of trials and utilized multiple representations of the simulated data made appropriate inferences regarding fairness. In addition, consulting and collaborating with a partner were important aspects of students' reasoning about the empirical results that contributed to their inferences and data-based decisions.

Making Connections Between Empirical and Theoretical Probability:
Students' Generation and Analysis of Data in a Technological Environment

Imagine playing your favorite board game where the number of moves is determined by rolling a standard six-sided die. Experienced game players often accept equally likely probabilities for each outcome. Moreover, observations of outcomes of repeated die rolls in game-playing experiences often support this assumption. Beginners do not have this experience to draw upon and thus initially infer theoretical ideas of likelihood based on the outcomes of playing games with dice. Consequently, students may initially develop notions regarding the theoretical probability of any given number that lead to particular generalizations (e.g., “I don’t get 6 a lot so 6 must be hard to get”, “Each number came up about the same amount, so this die seems fair”). Researchers (e.g., Green, 1983; Kerslake, 1974; Watson & Moritz, 2003) have found these types of generalizations evident in students’ belief about dice.

Borovcnik, Bentz, and Kapadia (1991) summarized differences between various ways to view probability. For this research, three of the views are relevant: (1) classical, (2) frequentist, and (3) subjectivist. The *classical* view regards probability as a quantity theoretically derived from the sample space, determined a priori trials being conducted. For example, from a classical perspective a game player assumes there is $1/6$ chance of landing on a 4, based on the equiprobability of the six outcomes in the sample space. In the *frequentist* view, probability is a quantity experimentally derived from data, determined a posteriori trials being conducted (e.g., a student’s conclusion of “ $1/6$

chance” to get a 4 based on results from game-playing experiences).¹ Both the classical and frequentist view are objective in the sense that one can use clear data-based arguments to support the estimated probabilities. In the *subjectivist* view, however, persons often update their probabilities based on a “learning from experience” model (Borovcnik, Bentz, & Kapadia, 1991, p. 42) and believe the probability of an event can change based on these experiences. For example, a subjectivist might recall instances where there were few outcomes of 4 and thus might infer that all outcomes are not equally likely. Sometimes students may also adjust their theory to account for new data. For example, they may conduct repeated trials and observe that frequencies of all outcomes are about equal. In later trials, if the outcome of 4 does not occur as often as other outcomes, a belief that rolling a 4 is less likely than other numbers might reemerge.

The key mathematical idea used to interpret empirical results in relation to theoretically-derived probabilities is the *law of large numbers*. This principle states that the probability of a large difference between the empirical probability and the theoretical probability limits to zero as more trials are collected. Thus, even after a large number of trials, it is possible to have an empirical probability that is substantially different than the theoretical probability.

There is general agreement that research on students’ probabilistic reasoning has been lacking sufficient study of students’ understanding of the connection between empirical and theoretical approaches to probability, particularly with the availability and use of simulation technology tools (e.g., Jones, in press). This particular deficiency is addressed in this study. Our research aims to examine how middle school students make

connections between experimental and theoretical probability when using technology tools, including various representations of data, to determine whether a simulated die toss experiment is fair. Because we are emphasizing data-based reasoning, we designed an instructional task to elicit students' use of a frequentist approach since they are not given a physical die to examine and are not asked to determine probabilities a priori. Within this context, we are specifically investigating the following questions:

1. How do students reason about the relationship between empirical and theoretical probability, the effect of sample size, and variability within and across samples of data obtained from simulations?
2. How does students' use of representations in the technology tool affect their analysis of sampled data and inform their decisions about fairness?

A brief review of literature will provide background for the constructs we use in our framework for analysis.

Related Literature

Several researchers have studied how students reason about probabilistic situations, both empirically and theoretically. Early research by Piaget and Inhelder's (1975) focused on children's development of theoretical probability concepts. They found that children's ability to compare probabilities improved as they transitioned from part-part to part-whole strategies and had stronger combinatoric skills. Fischbein (1975) was one of the first to posit that children possess intuitive understanding of relative frequencies and probability that can be mediated through instruction, even before strong part-whole and combinatoric strategies were well developed.

Tversky and Kahneman (1982) coined *representativeness* as a heuristic in which students tend to estimate the likelihood of an event based on how well empirical data represents some aspect of the population. For example, 10 die tosses that result in the outcomes of 1, 6, 2, 1, 3, 4, 2, 1, 1, 6 might be judged as less likely to happen than the results of 1, 3, 4, 2, 5, 4, 6, 1, 3, 5, 2 since the first set may be judged to have too many 1's and no 5's represented. However, as indicated by Lecoutre (1992), some students may have a natural tendency to use an equiprobable bias and associate chance and luck with events being equally likely by nature. In this case, they would judge all die toss results as equally likely since die have six equally likely sides and luck determines the outcome.

Another common heuristic appears when students interpret questions about the probability of an event to be a request to predict the result of a single trial. With such an *outcome approach* for making a prediction, students translate probability values into “yes/no” decisions (Konold, 1987; 1995). Thus, in comparing the two sets of die toss outcomes above students may interpret the question as determining whether or not each set is possible and answer “yes.” Students may also view the comparison as a request to determine whether the die is fair and answer “unfair” for the first set and “fair” for the second set, without regard to the context of the small sample size.

Several researchers have analyzed students' understanding and beliefs about the fairness of a standard six-sided die. Watson and Moritz (2003) found that elementary and middle school students hold strong beliefs regarding the fairness of dice and, in fact, generally doubt that each outcome is equally probable. In Green's (1983) study of more than 3000 early adolescents, he found that as age increases, the belief that six is “hard to

get” declines (23% for 11-year-olds down to 9% for 15-year-olds) and belief that all numbers on the die have the same chance increases (67% for 11-year-olds to 86% for 15-year-olds). In another study, Lidster, Pereira-Mendoza, Watson, and Collis (1995) found that some students believed a die could be fair: (1) for some numbers, (2) for some trials, or (3) compared to another die. These findings seem aligned with an “outcome approach” as described by Konold (1985; 1987). Such pervasive beliefs about dice are likely a product of students’ game-playing experiences and represent genuine challenges to mathematics teachers, particularly given the common use of dice (or number cubes) in popular mathematics curricular materials (e.g., Hake & Saxon, 2004; Lappan, Fey, Fitzgerald, Friel, & Phillips, 2004).

Researchers have also attended to students’ acceptance of variation in a sample of die tosses. When given expected results of 60 tosses of a fair die in three different bar graphs and asked to choose which one was most likely, 47% chose a graph with small variation, 36% chose the graph with large variation, and 17% choosing no variation (Green, 1983). Using a similar task, Lidster et al. (1995) showed third, sixth and ninth grade students graphs of 6 different dice, 3 fair and 3 unfair, for 60, 360, 600, and 12,000 trials. Most students were willing to believe a die was fair even with contrary visual and numerical evidence. Thus, they were willing to accept a wide variation in the sampled data that differed from expected for a fair die. Even though the results of tossing the die departed from what they expected or looked “strange,” even for graphical displays from large trials, students often concluded that the die was fair. The ninth graders in the Lidster et al. study seemed to have a more developed appreciation for the need for sampling large

trials when determining whether a die was fair based on the given graphical representations. However, in a recent study of third, sixth and ninth grade students, Watson & Moritz, 2003 reported that almost all students who believed die were fair or unfair did not use a strategy to collect data to confirm or refute their belief.

All of these studies concerning students' beliefs about the fairness of die allude to issues concerning sampling and sample size. Fischbein and Schnarch (1997) noted that acceptance in small sample sizes seemed to increase in older students in their study of 5th, 7th, 9th and 11th grade students. However, the tasks posed in their research asked students to compare results from different samples that were proportional (e.g., obtaining at least 2 heads out of 3 coin tosses versus obtaining at least 200 heads out of 300 coin tosses) to decide which event was more likely. Increasingly, the students seemed to choose the primary misconception that the events are equally likely and ignore the effect of the sample size in the task. Several researchers have noted that this misconception is quite likely due to an inappropriate generalization of proportional reasoning (Stavy & Tirosh, 2000; Van Dooren, Debock, Depaepe, Janssens, & Verschaffel, 2003). The contexts used by Lidster et al. (1995), Fischbein and Schnarch (1997), and Watson and Moritz (2003) were different and may allude to the effect of contextual reasoning on students' consideration of the effect of sample size. None of these studies asked students to determine an appropriate sample size and collect data themselves to determine the fairness of a die or to compare the probabilities of two events. We conjecture that a context in which students are asked to make inferences from experimental data may yield different reasoning about the effect of sample size.

Research on the role of simulation software in teaching and learning probability is a relatively recent endeavor, yet the National Council of Teachers of Mathematics (2000) encourages the use of technology-based simulations because they “afford students access to relatively large samples that can be generated quickly and modified easily” (p. 254). Recent research has begun to document what students do when they have access to technology tools for managing data from probability simulations. Drier (2000a, 2000b) found that fourth-grade students used the representations (e.g., bar graph, pie graph, data table) in probability simulation software as both objects to display and interpret data, and as dynamic objects of analysis *during experimentation* to develop a notion of an “evening-out” phenomenon. These students recognized that larger numbers of trials resulted in distributions that closely resembled what they expected from the theoretical probabilities used in the design of an experiment. Pratt (2000) and Pratt and Noss (2002) reported that 10-year-old students working with a different probability software environment developed an understanding of a connection between the number of trials and the distribution of data (viewed in pictographs and pie graphs). Pratt (2000) also reported that students used a “Workings Box” to control the sample space and theoretical probabilities, which they discovered affected the distribution of data. These research studies suggest that simulation tools that give students control over designing experiments, running as many trials as they desire, and viewing graphical representations of results may help in the development of deeper understandings of theoretical probability, empirical probability, sample size, and the relationship between them.

In a related study, Taylor (2001) found that the use of computer simulation software in whole-class instruction with one computer displayed for class discussions also has positive effects on upper elementary students' understanding of experimental probability. In particular, Taylor used one control group and three experimental groups to test the differences between using traditional lessons with no hands-on experimentation (control), computer simulations, concrete physical simulations, and a combination of computer and physical simulations. Although there was no significant difference between test results for the control and the groups using physical and physical/computer, there was a significant difference in students' improved understanding of experimental probability in the "computer only" group as compared to the control group. This significant difference was even more pronounced for students who had low pretest scores, leading Taylor to conclude that the whole-class use of displayed computer simulations were most beneficial for lower ability students.

Analytical Framework

Given the focus of our research questions, we adapted Stohl and Tarr's (2002) model for a bi-directional relationship between empirical and theoretical probability (see Figure 1). In this model, students reason about chance phenomenon either from or about theoretical probability. In the former case, students begin with beliefs regarding the theoretical probability of each event that can be from a classical or subjectivist approach. Thus, when rolling a die, students use their beliefs when interpreting empirical data with larger samples likely to be more representative of the theoretical probabilities while smaller samples may offer more variability and be less representative. In general, as the

sample size grows, relative frequencies of each outcome begin to closely approximate theoretical probabilities. In the other direction, students examine empirical data using frequentist or subjectivist approaches and use that data to inform decisions regarding the underlying (unknown) theoretical probabilities. For example, small samples offer little power in making decisions regarding fairness of a die; ten rolls may yield no 3's but such data hardly support the notion that rolling a three is an impossible event. By way of contrast, as sample size grows infinitely large, students should be able to more precisely estimate theoretical probabilities. It is this second type of reasoning – from empirical probability to theoretical probability – that forms the basis of inference making. Among the key statistical concepts students must attend to in this model, sample size is fundamental. However, in order to make sense of the notion of the law of large numbers, students need to coordinate conceptions of independent events and sampling variability on one hand, and the role of sample size on the other hand.

Although this study is aimed at making sense of students' probabilistic reasoning and is influenced heavily by research in that domain, we would be remiss not to recognize our underlying beliefs about student learning that guide our work. Our perspective acknowledges individuals' constructive process of resolving perturbations through reflecting on their actions (and subsequent effects) that allows for abstraction of ideas with a lens on the context in which meanings are socially negotiated through interactions (Tzur & Simon, 1999; Voigt, 1996; von Glasersfeld, 1995). Additionally, this perspective is augmented by the view that available tools and mathematical tasks may potentially enable and constrain learning (Wertsch, 1991; Graue & Walsh, 1998). Thus, students'

individual constructive process, available tools, mathematical tasks, and the social interaction and negotiations among students and between students and teacher will all operate interactively as students work on the probability task in this study.

To further specify important elements in our analysis, we employ the constructs of *external* and *internal resources* used in Pratt's (2000) study on students' probabilistic reasoning. Goldin (2003) also distinguishes between a learner's internal system of representation and systems that are external to the learner. This distinction allows for the recognition that one can not directly observe a learner's internal resource. However, by analyzing the interactions with external resources, we can make inferences about a learner's internal system. According to Noss and Hoyles (1996), students must coordinate these internal and external resources in a "webbing" (linking) process to construct meaningful knowledge. Thus, the notions of external and internal resources match well with our coordinated perspective of learning. External resources are those "that reside outside of the individual but within a setting" (Pratt, 2000, p. 605), and would include the social interactions, available tools, and the context of the mathematical task. Internal resources consist of an individual's understandings, including intuitions and beliefs. More details are given about how this framework is applied to our analysis in the methodology.

Methodology

Context of Study

The study took place during the first month of 6th grade (age 11-12) in an average-level mathematics class in an urban, public middle school in the southern United States. School demographics include 51% Caucasian students and 49% non-Caucasian students

with 29.5% of students receiving free or reduced lunch. The whole class (n=23) was engaged in a 12-day probability unit designed and taught by Stohl and Tarr. During the instructional sequence, students used real objects (coins, dice, spinners) and *Probability Explorer* (Stohl, 1999-2002) software on laptops to explore various tasks. The software used in our study was an updated version of the software used in research by Drier (2000a; 2000b) and Taylor (2001). Throughout the entire instructional sequence, small group tasks and whole class discussions were designed to elicit students' reasoning about connections between empirical and theoretical probability, and the role of sample size in making data-based inferences (see Stohl & Tarr, 2002 for details).

The Schoolopoly task (Figure 2) was the final performance assessment on days 10-12 of the unit. In the Schoolopoly task, students utilize the computer software to simulate rolling a die. The software enables users to set the number of trials, updates results after each trial, and allows students to view simulation data in iconic images on the screen, a pie graph (relative frequency), bar graph (frequency), and data table (frequency and relative frequency – fractions, decimals, and percents). Additionally, for each die company (see Figure 3), weights were assigned to each event, 1-6, but remained hidden from students during the activity.

Three Case Study Pairs

Prior to instruction, three case study pairs were selected based on scores on a standardized mathematics achievement test as well as a pretest on probability concepts developed by the researchers. With both pretests, we grouped the scores in thirds (high, middle, low). Candidate case study students were chosen as students who were relatively

consistent in their class ranking on both pretests (e.g., high on general mathematics test and upper range of middle or high on the probability test). Six students were then chosen to be collectively representative of gender and ethnicity of the class; the entire class (14 boys and 9 girls) consisted of 6 African American, 2 Hispanic, and 15 Caucasian students.

Dannie and Lara (2 Caucasian girls) represented the high-scoring group who investigated the Dice R' Us company for the Schoolopoly task. Brandon (Caucasian boy) and Manuel (Hispanic boy) represented an average-scoring group who investigated High Rollers, Inc. Greg (Caucasian boy) and Jasyn (African-American boy) comprised the low-scoring pair who investigated the Slice N' Dice company. See Figure 3 to examine the difficulty levels of the theoretical probability distributions for these three companies. All three companies produced biased dice, with Dice R' Us having a slightly better than expected chance of rolling a 6 (20% versus an expected 16.7% on a fair die) and slightly lower chance of rolling a one (13.3% versus an expected 16.7% on a fair die). This slightly biased die was given to Dannie and Lara in anticipation that they had a more highly developed appreciation for sample size and would likely develop the notion for the need to collect a large number of trials to detect bias in the die.

Sources of Data

All students in the class were seated in pairs or groups of three at tables in their regular classroom with a PC laptop, calculators, and manipulative materials (e.g., dice, spinners) readily available. The three case study pairs worked at three different tables where their laptop computers were connected to a PC-to-TV converter to video-record

their computer interactions while microphones captured their conversations. In addition, there was a video camera focused on the three tables to capture students' social interactions with each other and the teachers-researchers. For each pair, the videos were directly transcribed and annotated with screenshots and researcher notes to describe the actions of the students. For the Schoolopoly task, the students were routinely saving data and making notes during data collection in a word processing document. Each pair also constructed a poster with the answers to the three questions in Figure 2 and appropriate data displays as evidence to support their claims. The word processing documents and posters were also used in the analysis.

Enacting our Analytical Framework

The nature of the Schoolopoly task excluded the possibility of a classical a priori analysis of the theoretical probabilities based on symmetry of the die or numerical computation. Instead, a frequentist approach could have been used by students to make a posteriori decisions based on collecting and reasoning from data to make claims and to estimate (unknown) theoretical probabilities. Students were required to make a decision regarding the fairness of a die and support that decision with compelling evidence, thus providing a structure for social negotiations among partners. The task sought to foster students' appreciation for the role of sample size by purposefully not prescribing a sample size and with students determining how many trials were sufficient to support their conclusions. Finally, students had to decide how data should be represented iconically, numerically, and graphically as they analyzed data, as well as for supporting their claims and communicating their arguments on the poster.

Our intent was to analyze how pairs of students coordinated external and internal resources in their own coordination of part of the bi-directional relationship for how empirical probability or data distributions could be used to make claims about an unknown theoretical probability or distribution. Several external and internal resources were pertinent to this analysis.

Internal resources. Internally, students needed to bring forth understandings of independence, sample size, fairness, and variability. They should have recognized that every roll of a die was independent from the previous, as well as the ramifications of conducting independent sets of trials or ones that were dependent on another (i.e., were they pooling data where results from a set of 100 trials was added on to a prior set of 100 trials?). This type of decision became necessary as they considered and decided upon the appropriate sample size for each trial set.

The notion of fairness of a die was of concern in this study since students were explicitly asked to determine whether or not the die they were investigating was fair. Some students may have had an underlying subjective belief and assumption that all die are fair or have had a desire for their die to be fair. Students with an assumption of fairness may also not have needed to test a die with large number of trials (Watson & Moritz, 2003), or may have had an expectation of unpredictability (Pratt, 2000).

In reasoning about the variability of data, Mooney (2002) reported that students may compare data within a data set, or across independent data sets. Thus, it was important that we attended to both types of comparisons. We also wanted to infer from

students' work whether they were accepting of the variability in the data as compared to what they expected.

External resources. There were several external resources that were pertinent in our analysis. The first resource was the task itself (see Figure 2) and the classroom context in which the task was posed, including expectations set forth by the teacher-researchers. Within the context of this setting, students worked in pairs and were asked to come to consensus. Thus, the nature of the task encouraged social negotiations in order for students to convince themselves, persuade or reach consensus with their partner, and then be prepared to justify their claim to the whole class on their poster.

Students also had access to the software as an external resource and made decisions about which tools to use to represent the data during and after a simulation. The representations available included a stacked iconic pictogram, bar graph, pie graph, and a data table that could be viewed to include frequencies or relative frequencies. The students also were using general technology capabilities to copy, paste and save representations and whole screenshots in a word processing application to be used as evidence of the claims they were making.

An initial analysis of the videos and annotated transcriptions brought students' key decisions and actions in data collection and analysis to the fore. Thus, in order to do a fine-grained analysis of students' work, we broke each of the pair's work into "cycles" that began with their decision to run a set number of trials and ended with a decision to save the data as evidence for further analysis later and possible evidence in a claim, and whether or not to proceed to collect more data (see Figure 4).

For each question posed in Figure 4, we created codes based on our analytical framework to record students' decisions or actions within each cycle (see Table 1). In developing codes for sample size that could be consistently used across all three pairs of students, we considered the theoretical probabilities of the events 1-6 for the three companies: (1) Dice R Us, (2) High Rollers Inc, and (3) Slice-n-Dice (Figure 3).² The resulting five sample size levels (see Table 1) were then used for each pair. The coding process resulted in a condensed description of the important aspects of students' work that could help us quickly summarize each cycle. These coded cycles then allowed us to examine patterns within and across pairs of case study students.

Analysis

We discuss our analysis in two parts. First, we offer a description and analysis of each case study pair. In this analysis, we consider students' selection and use of representations available in the software, sample size, and social negotiations to infer the linking of their internal and external resources. Secondly, we examine the patterns that emerged across the three case study pairs.

Description of Case Pairs' Work on Task

Pair 1. Dannie and Lara employed 20 cycles with more than 50% having a cumulative sample size of 40 or below (Level 1). Approximately 73% of the cycles had a cumulative sample size of 100 or less. This group never used more than two representations at one time. Their reasoning was heavily influenced by their use of the pie graph, which they enabled in all cycles. During seven cycles they also used the stack representation and during five cycles they used the frequency table. The only other

representation utilized in the last cycle was the relative frequency table after they realized that a required element for their poster was to estimate the probabilities of each outcome. They did not use relative frequencies to offer any information about the fairness (or lack thereof) of the die.

Most of their cycles were similar, using Level 1 sample sizes, making the same conclusion of “pretty fair,” with the majority of their reasoning about variability occurring within a cycle, as they tended to use the pie graph to compare the relative amount of each outcome in relationship to each other. In the early cycles, they focused on the “evenness” of all outcomes except for the small number of 5’s. Dannie expressed interest in saving data sets where the results looked even, while Lara thought they should make decisions to save data “if it looks kind of even or if it looks really *not* [italics added] even.” Thus, early on, Dannie had a goal to show the die were fair while Lara had a goal to collect data to provide evidence of fair or unfair. These students also occasionally reasoned across the data sets. For example, in Cycle 11 Dannie noted that the company was fair for certain trials (e.g., the current set of 36 trials) and not fair for others, “because if we do six (trials) it is probably not going to be even.”

Although Dannie often argued that the data appeared even, the variation visible in the external representations was not convincing for Lara to wholly accept the classification of fair. Thus, they usually compromised to use the term “about even” or “pretty fair.” Dannie maintained her beliefs about die “eventually” being fair (i.e., an expectation that results *should eventually* “even out”) to justify the variation in the

external representation of the pie graph, even with small sample sizes. At one point, Lara reminded Dannie “We don't know if it's fair, because that was 12 [trials].”

A disagreement between the partners occurred in cycles and 18 and 19 when the sample size reached over 500 (Level 4). Dannie was on a quest “to see how long it takes before the die is fair.” One interpretation of her behavior is that she did not accept the variability in the outcomes of the die as indicating non-equiprobable events. Perhaps she expected that the fairness of die would become evident at some point in the future. The following dialogue occurred in Cycle 19 as they decided what to report about the outcomes for their trial of 906 rolls of the die.

Lara: ... Dannie do you want just paste it [in the document] and say that after 906 trials it wasn't fair?

Dannie: It looks *pretty* fair. [italics added]

Researcher: Is there any other tools that can tell you some more information?

Lara: Yeah.

Dannie: This [Dannie opens the Table and clicks in each of the count cells. 1-115, 2-139, 3-179, 4-172, 5-177, 6-124]

Lara: Okay.

Researcher: So what does that evidence tell you?

Dannie: It's sort of fair.

Researcher: It's sort of fair? So what is sort of *not* fair? [italics added]

Dannie: This ... This, these three [gestures with the mouse to the Table and points to the counts for 3, 4, & 5] the 3, 4, and 5.

Researcher: What about the 3, 4, and 5?

Dannie: They have more rolls than the 1, 2 and 6 ...

Lara: Okay, so we could copy that [pie] graph and say just what you just said.

[Following a brief exchange about copy/paste procedure to save whole screen of data.]

Lara: [reading as she types in document] We found that after –

Dannie: 906

Lara: [reading as she types] 906 trial, rolls, it was about fair.

Dannie: No it was fair [firmly], about fair doesn't really make sense, you know.

Lara: Yeah, but it isn't fair.

Dannie: [hesitantly] Okay.

Lara: It was almost perfect.

Dannie: Pretty.

Lara: Pretty fair.

Dannie's use of the external representations (data table and pie graph) indicated that although she pointed out distinct variations among the number of each outcome and initially stated the die was "pretty fair," this variation was not great enough for her to recognize a pattern in the data and she did not appear to consider the sample size in her eventual claim that the die was fair since "about fair doesn't really make sense." Her

expectation and acceptance of variability within and across data sets could have been due to her underlying belief that dice are fair or always “eventually get fair.”

The large sample size (906 trials) gave Lara confidence to argue for a position of “about fair” and, when Dannie wanted to claim fair, Lara objected with “but it isn’t fair.” This exchange indicated that Lara likely suspected that the die was not fair but was willing to agree to the claim of “pretty fair.” In this case, the external resources of the pie graph and data table, along with the large sample size, were useful for Lara’s understanding of how she could use the empirical data to make an inference about fairness. However the external resource of the social negotiations with Dannie were hindering Lara’s commitment to a claim of unfair and may have prevented her from establishing a stronger connection between the empirical results and the theoretical probability. Neither Dannie nor Lara made any reference to results from previous data sets in which the number of 1’s or 6’s was lower than the other outcomes. This may have also been due to their use of the pie graph, rather than the data table as an external representation of the data.

Throughout the activity, the girls’ attention was on providing evidence to make a determination of fair or unfair. It was when they were making their poster that they realized they had not estimated the probabilities of the outcome of the die. Lara returned to the software and conducted an additional cycle with a cumulative Level 1 sample size. From this sample, she used the relative frequency table and the empirical probabilities of $4/36$, $10/36$, $4/36$, $2/36$, $9/36$, and $7/36$ as estimates for the “probability after 36 trials” for each number on the die (see Figure 5). Her actions suggested that she had partitioned the

problem and was working on a separate task (estimating probabilities) without coordinating her reasoning on this task (estimating probabilities) with the claim of fairness about the die. Their estimate of unequal theoretical probabilities did not correspond with a conclusion that the die was fair (equally likely). Thus, according to the model in Figure 1, neither student made a meaningful connection between the decision of fairness and the reporting of unequal probabilities. Another possibility was that the students misinterpreted Question 3 in the task (see Figure 2) as a request to provide empirical probabilities. Perhaps, also, Lara decided to put the unequal probabilities on the poster as her way of indicating a possibility that the die was unfair.

This pair primarily used the pie graph representation on their poster with only one of the pie graphs from a Level 2 sample size (right-most pie graph on poster, see Figure 5). The remaining data displayed were from Level 1 cycles with cumulative sample sizes of 10, 20, 36, and 6. The image of the stacked pictograph and pie graph with 6 trials was the screenshot they took when the whole class was given directions about how to copy a screenshot and paste it into a word processing document. These students had forgotten how to copy and save a whole screen image during their data collection, which may explain their over-use of saving and using the pie graph image (although the same copy/paste operation could be used to save the bar graph or data table as a single representation). They also chose not to use the whole screen image (showing pie graph and frequency data table) of the data from Cycle 19 of 906 trials as evidence on their poster.

Pair 2. Brandon and Manuel conducted eight cycles, four of which had sample sizes of 100 or below. For all but two cycles, they had at least three representations visible: bar graph, pie chart, and frequency table. An exception was observed in Cycle 1 when they only used the stack representation and a subsequent cycle where they used the stack and the frequency table. For the final three cycles, they used the relative frequency table in addition to the other three representations listed. Both Brandon and Manuel referred to the different representations and justified their reasoning by using several representations and making connections among them. Their use of multiple representations was an essential component in detecting bias in their die and accurately estimating the probabilities of each outcome.

As in the case of Dannie and Lara, there was agreement between Brandon and Manuel that the die was fair for the early cycles with smaller sample sizes. However, once the sample size increased to over 100 (Level 3) disagreement began. Brandon was intolerant of the variation among the outcomes of the die within the trial and wanted to conclude that the die was not fair. Manuel, however, exhibited a high tolerance for variability within the data set and convinced Brandon that the die was fair by pointing out that “they all [outcomes] didn’t have to be the same...it was just the luck.” Brandon expressed his skepticism but did not try to convince Manuel of the unfairness of the die until the sample size reached above 1000 (Level 5). A key moment in the pair’s experimentation and dialogue is illustrated below. Following the transcript we describe our analysis of their use of internal and external resources and the connections they are making. In this context, the pie graph, bar graph and relative frequency table were

displayed. The students had set the software to initially run 1000 trials and they were at approximately 800 trials when the dialogue below began. Heretofore they have been arguing about fairness, Brandon claiming the die were unfair and Manuel claiming that it was fair because “they don’t all have to be even, man,” and stating, “I bet your wrong. I bet we're fair.”

Brandon: I bet we aren't fair.

Manuel: Well I don't care. We are fair. Just because it’s not all even doesn't mean we're not fair. [Clicks on RUN again to add an additional 500 trials.] Dude, were already up to a thousand [trials].

Brandon: It’s not. I really don't think...

Manuel: Well, I do. So...

Brandon: It's only beating it by about a hundred [referring to bar graph].

Manuel: It's not that unfair. See, 3 and 5 and 1 are practically the same.

Brandon: So they [3, 5, and 1] must have the same probability but that [points to relative frequencies for two and four in table] might have been more because look these [2, 4, and 6] are always a bit higher than these [1, 3, and 5]. And you could do two thousand trials [cumulative sample size currently at about 1500].

Manuel: [interrupting] Wait a second. Wait a second. We *are* unfair. These two, all these [1, 3, and 5] are one and these [2, 4, and 6] are two

[referring to hypothesized weights for each outcome as he points to the frequencies in the table]. So, 2, 4, 6... We're unfair.

During this cycle, Brandon was convinced the die was unfair and Manuel justified a claim of fair by noting that outcomes 1, 3, and 5 had similar relative frequencies. Not only did these students make connections among the external representations (particularly the bar graph and relative frequencies), but they each used those external representations to justify and make connections to their individual (and different) internal reasoning and conclusions. The social negotiations helped facilitate their connections among internal resources (sample size, variability, fairness) and an understanding of the relationship between empirical and theoretical probabilities. This is evidenced by Manuel's revelation that the die was unfair by Brandon's reference to the empirical distribution of data implying equal probabilities for outcomes 1, 3, and 5 and that these are less than the probabilities for 2, 4, and 6. Brandon's comment indicated that he was making a strong connection between empirical and theoretical probability. Brandon's verbalization of his internal understandings, along with the other external representations available that displayed the dynamically changing results as the sample size increased, may have influenced Manuel's ability to come to understand the connection between empirical and theoretical probability.

Brandon and Manuel's poster highlights the connections they made between their internal understanding of fairness and their external use of representations. It also displays the connections the students made among data from small and large independent samples, the representations, and their estimated theoretical probabilities as evidenced by

their arrows drawn on their poster. Brandon and Manual utilized the bar graph, pie graph, stacked pictograph, the frequency table, and relative frequency table during their simulations and used each of these representations to support their claim of fairness and their estimated probabilities. The estimated probabilities stated on the poster were rounded from the relative frequencies for 6000 trials (screenshot in middle of poster) and slightly adjusted to be integer percents that sum to one.

On their poster, they documented how they once asserted that the die was fair with a small sample and then how the larger samples gave them more confidence to make a claim of unfair and to estimate the probabilities. They included two screenshots from small sample sizes with captions stating “so far we need more evidence to conclude that the company may not be fair or [may be] fair” (6 trials) and “We have concluded from the info that we collected so far that the dice are so fare [sic]” (50 trials). They also included two screenshots from very large samples with the bar, pie and relative frequencies displayed. For 2000 trials they noted, “We have changed our minds, the dice are sooooo sooooo sooooo [sic] unfair. We think that cause [sic] this is our hypothesis” and then for 6000 trials, “We think that it is still unfair because of the evindecne [sic] of the percent on the table.” They also noted that, “The info we collected later proved that high rollers was unfair! 6, 4, 2 had higher probability then [sic] 5, 3, and 1.” Their use of a variety of sample sizes and their statements indicated that they were able to coordinate the concepts of sample size, variation, and independence while making strong connections between empirical results and underlying theoretical probabilities.

Pair 3. Greg and Jasyn employed 11 cycles with sample sizes ranging from six Level 1 samples (40 or less) to a single Level 4 sample (501 to 1000). Their use of representations ranged from a low of two representations on two cycles (pie graph and stack on Cycle 1, and pie graph and frequency table on Cycle 5) to three representations, most commonly the bar graph, pie graph, and frequency table. Greg and Jasyn never utilized the relative frequency table, which might have contributed to their inability to estimate the probability of the six outcomes on the die.

Because of the relatively large variation in the theoretical probabilities for Slice and Dice (see Figure 3), they did not need to take extremely large samples in order to confidently make inferences regarding fairness. In Cycle 2, they conducted a Level 4 sample with the bar graph, pie graph, and frequency table open. As they began this cycle, they previously had 61 trials, then ran 500 added on to this data set. The following conversation began during the simulation at about 264 trials with a low number of fives.

Jasyn: Dang! Look at [the event] 5, there is only 7 [referring to the frequency table]

Greg: So this dice is unfair.

Jasyn: Our dice is very unfair, there is only eleven 5's.

Greg: Okey, dokey, we should get five hundred and ... twelve [referring to frequency of 5s in table], oh my oh my.

Jasyn: Let's paste this graph once it is done [pointing to Pie].

Greg: No problem.

Researcher: Whoa, what's going on with yours? [at 430 trials]

- Jasyn: This dice is probably very unfair, because–
- Greg: There is a very limited amount of 5's.
- Jasyn: It's 5 and it's only 20 [referring to frequency of 5's in table] the rest are like way above.
- Greg: The rest are like a hundred. The other ones are up there. 1's a little low too [pointing to count of 1 in Table] and ah, and ah...
- Jasyn: No, 1's not low [simulation now stopped, see Figure 7].
- Greg: 2 is high.
- Jasyn: 5 is way too low to be a fair dice.
- Greg: Copy. [saved Pie Graph in word processing document and recorded "This graph shows 561 trials. This [company] seems unfair. 5 is not coming up a very little amount."]

Their work in this cycle enabled them to observe the extremely low number of outcomes of a roll of 5 in comparison to the other die outcomes. Their coordination of external representations with internal decisions and intuitions of sample size and tolerance of variability further enabled Greg and Jasyn to make an early inference that their die was not fair. In addition, although Greg noted that the variability between 1 and the other numbers was questionable ("1's a little low, too"), Jasyn disagreed and they never revisited the notion of the likelihood of 1 being too low. Greg also pointed out the frequency of 2 being a bit high, which indicates he may have had a low tolerance for any event differing from exactly even (as the frequencies for 3, 4, and 6, are very close). However, he never made subsequent claims about 2 as compared to the other outcomes.

Following this cycle, the remaining cycles had relatively small sample sizes. Their results supported their previous finding of too few 5's and they concluded that the die was biased. In essence, their one large sample size seemed to be enough evidence for them to make their conclusion. This was an artifact of an external resource of the task design for their company (Slice and Dice) and most likely contributed to the lack of a need for much social negotiation between the two students.

Greg and Jasyn struggled with the general technology of copying and pasting their evidence into a word processing program. This, combined with their lack of use of the relative frequency table, hindered their ability to attend to the task of estimating the probability of the outcomes of the die. As shown on their poster (Figure 8), they used evidence from a Level 2 sample size rather than the Level 4 sample illustrated in the dialogue above to justify their claim, even though they had saved the pie graph with 561 trials. In addition, they did not answer the question regarding an estimate for the probabilities. They appeared to only be focused on providing evidence to convince others that 5 did not appear much in their experimental data.

The external resources of the task design for this pair (e.g., skewed bias in Slice-n-Dice probabilities) combined with their use of a large sample size in Cycle 2 fostered a coordination of internal and external resources that enabled the pair to quickly reach a conclusion about the fairness of the dice produced by their company. In particular, they were able to coordinate their internal ideas of the expected variability of outcomes of fair dice with the external representations of the bar graph, pie graph, and frequency table. This coordination allowed them to conclude that the low number of fives visible in the

representations was not acceptable in fair dice. They also used the external resources to justify to the teacher-researcher why their dice could not be fair, thus demonstrating their networking of internal understanding and external representations.

Greg and Jasyn's poster (Figure 8) illustrates their use of the pie graph, bar graph, and frequency table. These were the representations they utilized during their simulation and to justify their claim of unfair. They offered frequencies but did not seem to realize that theoretical probabilities can be estimated using relative frequencies. Their lack of use of the relative frequency table may have been a contributing factor to their inability to estimate the theoretical probabilities.

Patterns Across Case Study Pairs

When analyzing across the three case study pairs, consideration was given to use of both internal and external resources. This analysis first examined how students, across the cases, grappled with internal issues of sample size, fairness, variability, and independence. External resources such as the use of representations available in the software, use of technology to copy and save into a word processor, task context, and social negotiations are then discussed.

Internal resources: Sample size, fairness, variability, and independence.

Consistent with the findings of Lidster et al. (1995), some students believed that the dice could be fair for "some numbers" and for "some trials". In almost every cycle across the three case study pairs, students typically rendered a judgment of fairness during or after a set number of trials were carried out, with each group having at least one instance of claiming "fair" and "unfair" during a particular cycle. This phenomenon may be an

artifact of the context of the task asking for a judgment of fairness, although the task asked for a final judgment and did not request that students make intermediate claims during data collection. Some students also demonstrated a strong desire for their die to be fair, possibly because of an underlying assumption of fairness, or the social context of hoping the company they were examining would be the “winner” and chosen to supply dice for the Schoolopoly game. In contrast to the findings of Lidster et al. (1995), the majority of students in this study tended to overcome these beliefs as they obtained larger sample sizes and engaged in discourse with their partner. Students who ran large numbers of trials made more substantive inferences about the fairness of their die (Brandon and Manuel, Greg and Jasy).

When reasoning within or across data sets, students exhibited varying levels of tolerance of variation. What they individually considered to be a “normal” (expected) or unacceptable amount of variation was related to different patterns they observed in data. For example, having an expectation of unpredictability (e.g., Manuel’s sense of “it’s just the luck”) or that all dice are eventually fair (e.g., Dannie) contributed to students’ expectation and acceptance of a large amount of variance from even across and within the data sets.

Within cycles of larger sample sizes, substantial disagreements between students often occurred. These conflicts generally arose when one student was willing to accept a large amount of variation between outcomes for a fair die or across sets of trials while the other was unwilling to tolerate such a large degree of variation and therefore claimed the die was unfair. Of the two pairs who experienced conflicting opinions about their die, one

pair (Brandon and Manuel) was able to reach consensus regarding fairness after their sample size became large by using the data to make arguments and justify their claim. By way of contrast, the other pair (Dannie and Lara) agreed upon language (e.g., “pretty fair”) that reflected the relative non-commitment by Lara to accept Dannie’s claim of fairness. However, neither student made strong data-based arguments to support each other’s claim. One contributing factor to this unresolved issue may have been their use of the pie graph coupled with a limited use of the data table. The pair (Greg and Jasyn) analyzing the most blatantly biased die never conflicted on the issue of fairness. They did make some claims of unfair when they only used Level 1 sample sizes, but only after they had run several prior Level 1, Level 2 (41-100 trials), and Level 4 (501-1000 trials) cycles.

All groups utilized some independent runs of trials and some runs that were dependent on (i.e., combined with) previous sets of trials. Many of the dependent runs were used by students in the middle of a cycle when they decided, before the simulation stopped, to increase the number of trials. The most prevalent use of dependent trials was by Dannie who often added trials on to previous results as she kept expecting the die to “eventually be fair,” even though Lara seemed to prefer running several independent sets of trials.

External resources: Representations, technology, task context, and social negotiations. Groups who used fewer representations throughout their analysis were more prone to make faulty inferences regarding the fairness of the die. Across the groups there was a total of 39 cycles. Within some cycles there were instances when more than one

claim code (F, M, U, N, see Table 1) was recorded. This occurred when the students disagreed and expressed conflicting claims or when the students would make claims at more than one point during the data collection and analysis cycle (e.g., claiming “sort of fair” after 40 trials but by 200 trials claiming the die was “fair”). There were also ten cycles when none of the students made any verbal claim as to the fairness of the die. Table 2 reports the counts for each claim made across the three pairs when students had 2 or fewer or more than 3 representations available when the claim was made. We combined the claim code for “moderately fair” (M) and “fair” (F) because we consider these as instances when a student verbalized a claim other than the correct claim that the die was unfair.

It is important to note that 17 of the 18 claims of “moderately fair” or “fair” made with 2 or fewer representations were made by either Dannie or Lara whose claims were based on a pie graph as one of their representations. However, when three or more representations were available, six of the unfair claims were made by Greg and Jasyn, while all five of the moderately fair or fair claims were made by Brandon and Manuel. Although these initial results can hint at how the use of few representations and a reliance on the pie graph may yield incorrect claims, these results need to be tested under a more controlled experimental context with a larger sample of students.

All case study pairs started out using the stacked pictograph representation. However, a desire to use the stacked pictograph may have contributed to some students desire to run smaller samples. The students realized earlier in the unit of study that because of limited screen space in the software, larger trials tended to result in the number

of stacked data icons exceeding the height limit and then remaining data icons scattering randomly across the screen. Although no group verbalized this explicitly during their work on the Schoolopoly task, they appeared frustrated when the icons would scatter and would subsequently use a smaller sample size in the next cycle.

The pair (Brandon and Manuel) who was the most successful in both determining the die was unfair and making an estimate of the theoretical probabilities of the outcomes of the die used the frequency table after the run much earlier than the other two groups. They were also the only group to attend to the relative frequency table both during and after the run, with the exception of Dannie and Lara who used it on their final run to determine probabilities for inclusion on their poster. They did not use it to determine fairness or unfairness and, in fact, did not correctly determine that the outcomes on their die were not equiprobable. The group (Brandon and Manuel) who used the relative frequency table during and after the run referred to these values often in their communication with each other and with the teacher-researcher to justify their claims of fair or unfair.

Some students also displayed idiosyncratic thinking (Mooney, 2002) when they classified the die as “about even.” This may have been attributed to their use of the pie graph representation, which made it difficult to visually detect relative differences between pie slices. Furthermore, the bar graph may have highlighted the differences in each numerical outcome of the rolls of the die, whereas variations between relative frequencies were difficult to observe on the pie graph because it provided no direct

numerical information. These factors might account for Dannie and Lara's use of the terms "about even" and "pretty even" in their descriptions of experimental data.

It is interesting to note that all case study pairs generally compared outcome-to-outcome (e.g., "Five is beating one") rather than outcome-to-expected theoretical result. The availability of the table and bar graph seemed to facilitate these outcome-to-outcome comparisons. On several occasions, students used terminology that implied they were viewing the simulation as a "race" between the outcomes of the die (e.g. "Come on 5, catch up!", "Five is a bit behind"). The dynamic nature of the technology may have contributed to this phenomenon. However, this "cheering" phenomenon can also encourage "students to notice and interpret representations that provide new insight about the phenomena" (Enyedy, 2003, p. 378) and serve as a catalyst for focusing students' attention on variability within a run of trials.

The copying and pasting of data displays temporarily constrained students and seemed to influence which representations they saved as evidence. All case study pairs needed to be reminded of how to copy the whole screen and paste it into the word processing program. For Brandon and Manual, this reminder came early on allowing them to avoid some of the technological difficulties faced by the other two pairs. Greg and Jasyn attempted to paste a bar chart and a pie graph and when the two representations did not wrap around the page correctly, they used a significant amount of time trying to format the word processing document. This attention to the cutting and pasting activity appeared to redirect their attention off the task of estimating probabilities and may have

been a factor in their inability to complete that task. Dannie and Lara only pasted one graph (pie graph) at a time until they were reminded how to copy the whole screen.

Another technology factor observed across the three pairs of students was the issue of who controlled the mouse. When the partner who was not in control of the mouse wanted to use a different representation or run a different number of trials, the inclination was to demand the mouse (e.g., “let me do it”) or tell the controller what to do (e.g., “Come on, open the pie graph!”). Thus, students often had to verbalize and justify what they wanted to do in the software rather than just doing it themselves. In this way, the context of the shared computer further promoted student discourse and social negotiation.

The discourse fostered by the context of the task encouraged students to come to consensus about what they were attempting to do during each cycle and what they wanted to save as evidence after the cycle was over. We also encouraged students to reach consensus to determine whether or not each company produced fair die. This consensus building was established through social negotiations between the partners and could be observed in all pairs at some point. In the case of Dannie and Lara, they concluded the die was fair, but not by reaching a clear consensus. Ultimately, they compromised through use of phrases such as “about even” and “pretty fair.”

Discussion and Implications

The Schoolopoly task was intended to foster students’ reasoning from a frequentist perspective, using empirical data to make inferences about underlying theoretical or unknown probabilities. Students’ linking of their internal resources (sample size, independence, fairness, and variability) to external resources (representations,

technology, task context, and social negotiations) was critical in their ability to complete all aspects of the task (see Figure 2). Students who ran larger numbers of trials, utilized multiple representations, and negotiated with their partner were ultimately more successful in coordinating these internal and external resources and subsequently more accurate about their inference of fairness and their estimation of probabilities.

All students in this study had several prior experiences using the software in which they used the Weight Tool to examine, and in some cases to enter, the theoretical probabilities for various contexts (e.g., spinners, coins, marble bags). Thus, the Schoolopoly task and our analytic framework were explicitly designed from a frequentist perspective, with an expectation that students would make data-based arguments to support their inferences and estimations of underlying probabilities programmed into the software—probabilities known only to the teacher-researchers. Two pairs of students made appropriate data-based inferences about fairness with Brandon and Manuel being highly successful in developing an estimate of the underlying probabilities. The lower-scoring pair of students (Greg and Jasyn) may have struggled in making an estimate of the probabilities for a variety of reasons including technology difficulties, lack of attention to final question in task, difficulty in understanding how empirical evidence can be used to make estimated probabilities, and limited skills with fractions and percents. Nevertheless, this pair showed evidence of using data and various representations to make their claim of an unfair die.

The high-scoring pair of students (Dannie and Lara) was surprisingly the least successful in detecting bias and estimating underlying probabilities. Recall that the die

they were investigating was only slightly biased, making it more challenging to detect bias. Although we anticipated they would be better prepared to use large samples and make appropriate inferences, asking them to detect the slight bias may have been beyond their grasp. Their use of small sample sizes, over-reliance on the least useful (and potentially misleading) representation for detecting small relative differences (pie graph), and lack of co-constructing meaning between partners appear to be strong factors in their lack of success. However, this perceived lack of success was within the context of the use of a task designed to fit the researchers' frequentist perspective and knowledge of the underlying probabilities programmed into the software. How would the students' perceived intention of the task and researchers' analysis change if the task context were such that underlying probabilities were unknown to students and teacher-researchers? In such a context, a broader perspective in the analysis may be needed to include subjectivity.

One student (Dannie) demonstrated a primarily subjectivist approach to reasoning about probability evidenced by the way in which she updated her beliefs based on experience. For example, she claimed some sets of trials were "pretty fair" while others were not and she seemed to expect that if she ran enough trials the die would eventually become fair. Her reasoning could be considered as consistent with the bi-directional model (Figure 1) from a subjectivist viewpoint, rather than an objectivist view that utilizes classical and frequentist approaches. Some researchers have purported that a subjectivist approach to probability is more natural to children and may be a more sound

philosophical approach to teaching probability (Hawkins & Kapadia, 1984). This hypothesis, of course, needs to be further investigated.

Our research has helped to answer the call by Jones (in press) for more research on students' understanding about the connection between empirical and theoretical probability and the use of technology tools in learning probability. Given an engaging task and open-ended software tools, some students in this study were able to reason from a frequentist perspective and make meaningful connections among their internal resources and external resources. Artifacts such as the task context and data representations seemed to encourage students to make data-based arguments to support their claims to peers and teachers; and thus, played a critical role in fostering students' understanding of the relationship between empirical and theoretical probability. Although we used social negotiation as an external resource in our analysis, we did not focus on the intricacies of this negotiation and collaboration process. A follow-up analysis with a focus on argumentation and social interaction would further illuminate how students' interactions with each other affected their understandings.

We would like to offer a final word about the value of a frequentist perspective to probability in instruction. In most cases, everyday probabilistic situations in our world do not allow for a classical approach to computing probability. The medical field is a prime example of the use of data to estimate the likelihood of various phenomena (e.g., contracting a disease or developing complications during surgery). These probabilities are not theoretically-derived constructs based on a known sample space and (hopefully) not subjective opinions based on intuitions, but instead are calculated from large samples of

available data. There is a need for students to experience instructional techniques that foster an understanding of how such probabilities are determined, and that they represent an estimate of underlying probabilities and conditions that are unknown to us as external observers.

Granted, the Schoolopoly task is contrived and does not have the same real world importance as medicine; however, we found the problem context to be engaging to students and one that yielded rich classroom discourse. We ponder whether the context of dice as a familiar object in students' game-playing experiences added too much complexity in students' ability to reason about empirical and theoretical probability. It would be beneficial to study students' work if given a parallel task set within a different context that is not so laden with preconceived notions and game-playing experiences. Coupled with the use of simulation technology, problem tasks similar to Schoolopoly may 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. Thus, we believe that giving students access to appropriate external resources can help them make meaningful links among their internal resources of variability, independence, and sample size. A robust understanding of these three concepts may help students to develop an understanding of a bi-directional relationship between empirical and theoretical probability.

Notes:

¹ One cannot precisely determine the actual probability of rolling a 4 on a regular six-sided die because of the complex physics involved (e.g., the speed and angle at which the die is thrown, the initial spin of the die, air resistance – see Wolfram, 2002). However, we can use a classical Laplacean approach to embody the complexities of the physics and apparent (and probably imperfect) symmetry of the die and assume the probability of rolling a 4 is $1/6$. This theoretically derived probability of $1/6$ is an estimate of the actual probability that is unknown to us. If one rolls a die a given number of times, a frequentist approach can be used to state the experimental probability in terms of the proportion of 4's. But again, this experimental estimate only describes the probability of getting a 4 based on that set of random rolls. A repeated set of die rolls would most likely yield a different experimental estimate of the actual probability.

² In order to make a determination of fairness, the proportion of a certain outcome obtained by sampling must be compared to the theoretical population proportion of achieving that same outcome. Assuming, until proven otherwise, that the dice is fair, the theoretical proportion for any one outcome is $p = 1/6$. The margin of error for an inference about the proportion of any one particular outcome of the dice is calculated by the formula:

$$Error = Z_{\alpha/2} * \sqrt{\frac{p(1-p)}{n}}$$

To determine the sample size, n , for a desired margin of error, requires solving the formula for n , thus,

$$n = \frac{\left(Z_{\alpha/2}\right)^2 * p * (1-p)}{Error^2}$$

The outcome whose probability varies the most from the theorized probability will be the easiest to distinguish. The difference between that probability which is .04 (see Slice-n-Dice in Figure 3) and the theorized probability is $.167 - .04 = .127$. We chose .127 as our margin of error for our first level of sample size, because in order to confidently detect a difference from .167, the proportion in the sample would have to be within .127 of the theorized probability. For a desired confidence level of 98%, $Z_{\alpha/2} = 2.33$. The sample size is:

$$n = \frac{(2.33)^2 * \frac{1}{6} * \left(1 - \frac{1}{6}\right)}{.127^2} = 46.7 \approx 47.$$

We chose a sample size of 40 for the maximum in a Level 1 sample because we wanted a sample size such that no group should have been able to determine if their company was producing fair die. The margin of error for a sample size of 40 with a confidence level of 98% is .137.

Similarly, to choose our next level we considered the next most extreme (to 1/6) probability, .25 (see Figure 3). The difference between that probability and the theorized probability of 1/6 is $.26 - .167 = .083$. Thus we chose a margin of error of .083. Again with a confidence level of 98%, the sample size is:

$$n = \frac{(2.33)^2 * \frac{1}{6} * \left(1 - \frac{1}{6}\right)}{.083^2} = 109.45 \approx 110.$$

We chose a sample size of 100 for a Level 2 sample because we wanted a sample size in which only a minority of groups would be able to predict the fairness of the dice with confidence. The margin of error for a sample size of 100 with a confidence level of 98% is .086.

Continuing, difference between the next most extreme probability (.133) and the theorized probability of 1/6 is $\approx .034$, which produces a sample size of ≈ 654 . The margin of error for a sample size of 500 with a confidence level of 98% is .039. Thus choosing a sample of 500 or lower allowed some groups to predict fairness of the dice with confidence and some could not.

References

- Borovcnik, M., Bentz, H.-J., & Kapadia, R. (1991). A probabilistic perspective. In R. Kapadia & M. Borovcnik (Eds.), *Chance encounters: Probability in education* (pp. 27-71). Boston: Kluwer Academic Publishers.
- Drier, H. S. (2000a). *Children's probabilistic reasoning with a computer microworld*. Unpublished doctoral dissertation. University of Virginia.
- Drier, H. S. (2000b). Children's meaning-making activity with dynamic multiple representations in a probability microworld. In M. Fernandez (Ed.), *Proceedings of the Twenty-Second Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, (pp.691-696). Columbus, OH: ERIC Clearinghouse.
- Enyedy, J. (2003). Knowledge construction and collective practice: At the intersection of learning, talk, and social configurations in a computer-mediated mathematics classroom. *The Journal of the Learning Sciences* 12(3), 361-407.
- Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Boston: D. Reidel Publishing Company.
- Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic intuitively based misconceptions. *Journal of Research in Mathematics Education*, 28 (1), 96–105.
- Goldin, G. A. (2003). Representation in school mathematics: A unifying research perspective. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research*

- companion to principles and standards for school mathematics* (pp. 275-285).
Reston, VA: National Council of Teachers of Mathematics.
- Graue, M. E., & Walsh, D. J. (1998). *Studying children in context: Theories, methods, and ethics*. Thousand Oaks, CA: Sage Publications.
- Green, D. R. (1983). Shaking a six. *Mathematics in School*, 29-32.
- Hake, S., & Saxon, J. (2004). *Saxon math 7/6 (4th Edition)*. Norman, OK: Saxon Publishers.
- Hawkins, A. S., & Kapadia, R. (1984). Children's conception of probability – A psychological and pedagogical review. *Educational Studies in Mathematics*, 15, 349-377.
- Jones, G. A. (in press). Reflections. In G. A. Jones (Ed.) *Exploring probability in school: Challenges for teaching and learning*, (pp. 375-379). Netherlands: Kluwer Academic Publishers.
- Kerslake, D. (1974). Some children's views on probability. *Mathematics in School* 3(4), 22.
- Konold, C. (1987). *Informal concepts of probability* (Paper based on a Ph.D. Thesis.). Massachusetts: University of Massachusetts.
- Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education* 3(1), [online serial, available at www.amstats.org/publications/jse/v3n1/konold.html].
- Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (2004). *Connected mathematics grade 6*. Needham Heights, MA: Pearson Prentice Hall.

- Lecoutre, M. (1992). Cognitive models and problem spaces in “purely random” situations. *Educational Studies in Mathematics*, 23, 557-568.
- Lidster, S. T., Pereira-Mendoza, L., Watson, J. M., & Collis, K. F. (1995). *What’s fair for grade 6?* Paper presented at the annual conference of the Australian Association for Research in Education, Hobart, November.
- Metz, K. E. (1999). Why sampling works or why it can’t: ideas of young children engaged in research of their own design. In: F. Hitt, & M. Santos (Eds.), *Proceedings of the twenty-first annual meeting of the North American chapter of the international group for the psychology of education* (pp. 492–498). Columbus, OH: ERIC Clearinghouse.
- Mooney, E. S. (2002). A framework for characterizing middle school students’ statistical thinking. *Mathematical Thinking and Learning*, 4(1), 23-63.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- Noss, R. & Hoyles, C. (1996) *Windows on mathematical meanings: Learning cultures and computers*. Dordrecht: Kluwer.
- Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York: Norton.
- Pratt, D. (2000). Making sense of the total of two dice. *Journal of Research in Mathematics Education*, 31, 602-625.
- Pratt, D., & Noss, R. (2002). The micro-evolution of mathematical knowledge: The case of randomness. *Journal of the Learning Sciences*, 11(4), 453-488

- Stavy, R., & Tirosh, D. (2000). *How students (mis-)understand science and mathematics: Intuitive rules*. New York: Teachers College Press.
- Stohl, H. (1999–2002). *Probability Explorer*. Software application distributed by author at <http://www.probexplorer.com>.
- Stohl, H., & Tarr, J. E. (2002). Developing notions of inference with probability simulation tools. *Journal of Mathematical Behavior* 21(3), 319-337.
- Taylor, F. M. (2001). *Effectiveness of concrete and computer simulated manipulatives on elementary students' learning skills and concepts in experimental probability*. Unpublished doctoral dissertation, University of Florida.
- Tversky, A., & Kahneman, D. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 3-22). New York: Cambridge University Press.
- Tzur, R., & Simon, M. A. (1999). Postulating relationships between stages of knowing and types of tasks in mathematics teaching: A constructivist perspective. In F. Hitt & M. Santos (Eds.), *Proceedings of the Twenty First Annual Meeting of the North American Chapter of the International Group for the Psychology of Education* (pp. 805-810). Columbus, OH: ERIC Clearinghouse.
- Voigt, J. (1996). Negotiation of mathematical meaning in classroom processes: Social interaction and learning mathematics. In L. P. Steffe, P. Nesher, P. Cobb, G. A. Goldin, & B. Greer (Eds.), *Theories of mathematical learning* (pp. 21-50). Mahwah, NJ: Lawrence Erlbaum Associates.

- Van Dooren, W., De Bock, D., Depaepe, F., Janssens, D., & Verschaffel, L. (2003). The illusion of linearity: Expanding the evidence towards probabilistic reasoning *Educational Studies in Mathematics* 53(2), 113-138.
- von Glasersfeld, E. (1995). Sensory experience, abstraction, and teaching. In L. P. Steffe & J. Gale (Eds.), *Constructivism in education* (pp.369-383). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Watson, J. M., & Moritz, J. M. (2003). Fairness of dice: A longitudinal study of students' beliefs and strategies for making judgments. *Journal for Research in Mathematics Education* 34, 270-304.
- Wolfram, S. (2002). *A new kind of science*. Champaign, IL: Wolfram media, Inc.
- Wertsch, J. V. (1991). *Voices of the mind: A sociocultural approach to mediated action*. Cambridge, MA: Harvard University Press.

Figures

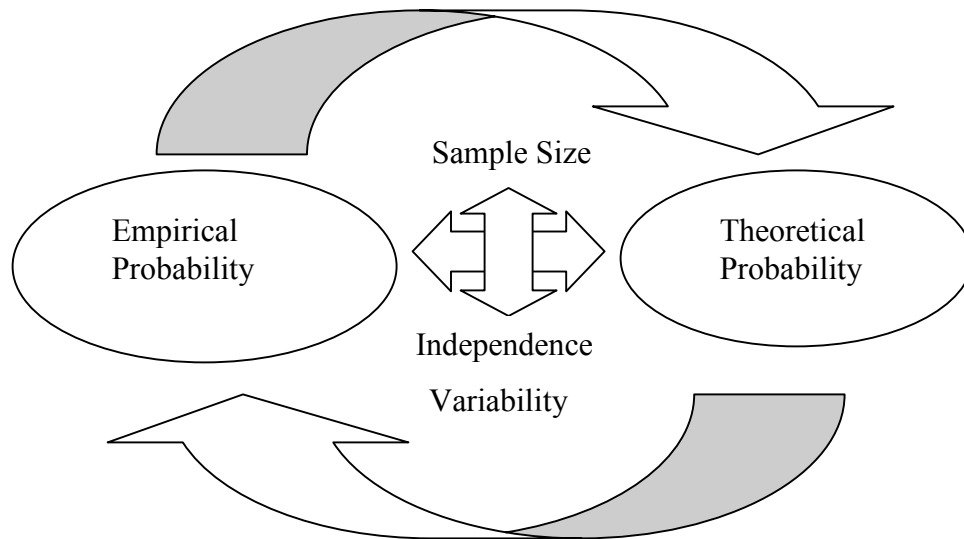


Figure 1. Bi-directional model

Schoolopoly

Your school is planning to create a board game modeled on the classic game of *Monopoly*TM. The game is to be called *Schoolopoly* and, like *Monopoly*TM, 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 a sample die for analysis and you will be assigned one company to investigate:

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

Your Assignment

Working with your partner, investigate whether the die sent to you by the company is, in fact, fair. That is, are all six outcomes equally likely to occur? You will need to create a poster to present to the School Board. The following three questions should be answered on your poster:

1. Would you recommend that dice be purchased from the company you investigated?
2. What evidence do you have that the die you tested is fair or unfair?
3. Use your experimental results to estimate the theoretical probability of each outcome, 1-6, of the die you tested.

Use Probability Explorer to collect data from simulated rolls of the die. Copy any graphs and screen shots you want to use as evidence and paste them in a Word document. Later, you will be able to print these.

Figure 2. Schoolopoly task as given to students.

Company Name	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	4 [0.133]	5 [0.167]	5 [0.167]	5 [0.167]	5 [0.167]	6 [0.2]
High Rollers, Inc.	2 [0.133]	3 [0.2]	2 [0.133]	3 [0.2]	2 [0.133]	3 [0.2]
Dice, Dice, Baby!	4 [0.133]	5 [0.167]	6 [0.2]	6 [0.2]	5 [0.167]	4 [0.133]
Pips and Dots	1 [0.167]	1 [0.167]	1 [0.167]	1 [0.167]	1 [0.167]	1 [0.167]
Slice n' Dice	4 [0.16]	5 [0.2]	5 [0.2]	5 [0.2]	1 [0.04]	5 [0.2]

Figure 3. Weights and theoretical probabilities for events 1-6 in each company.

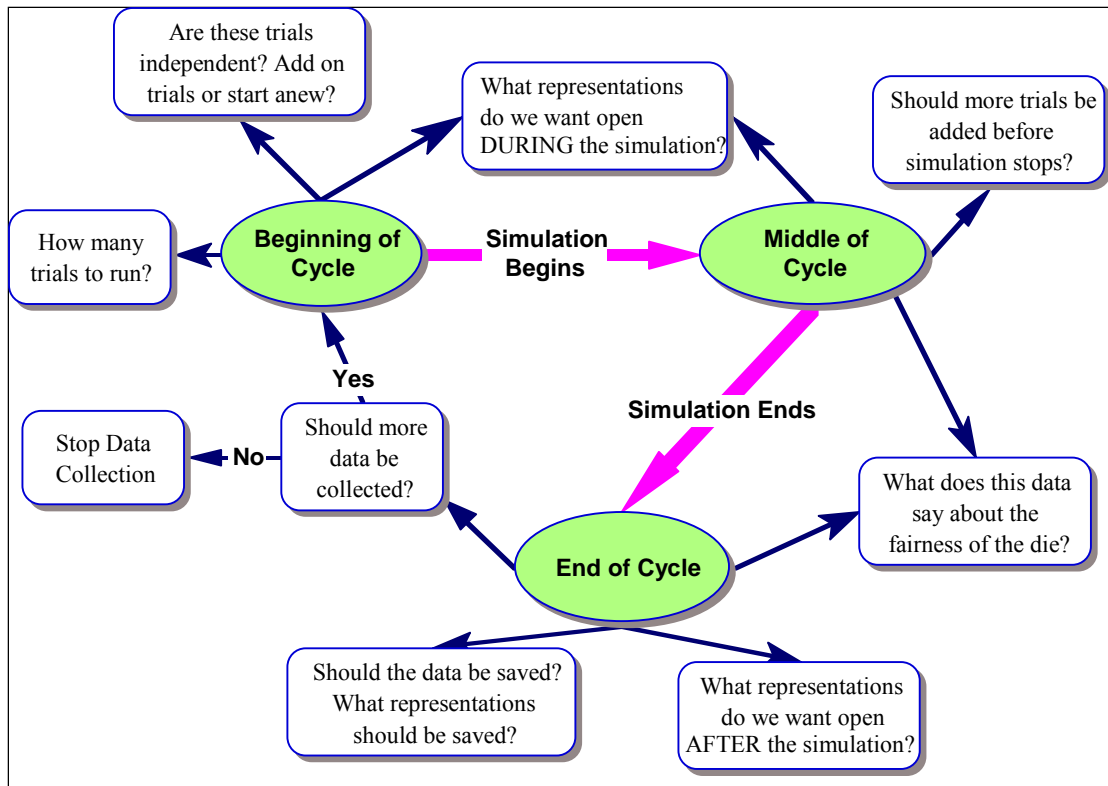


Figure 4. Detailed decisions students make during a cycle.

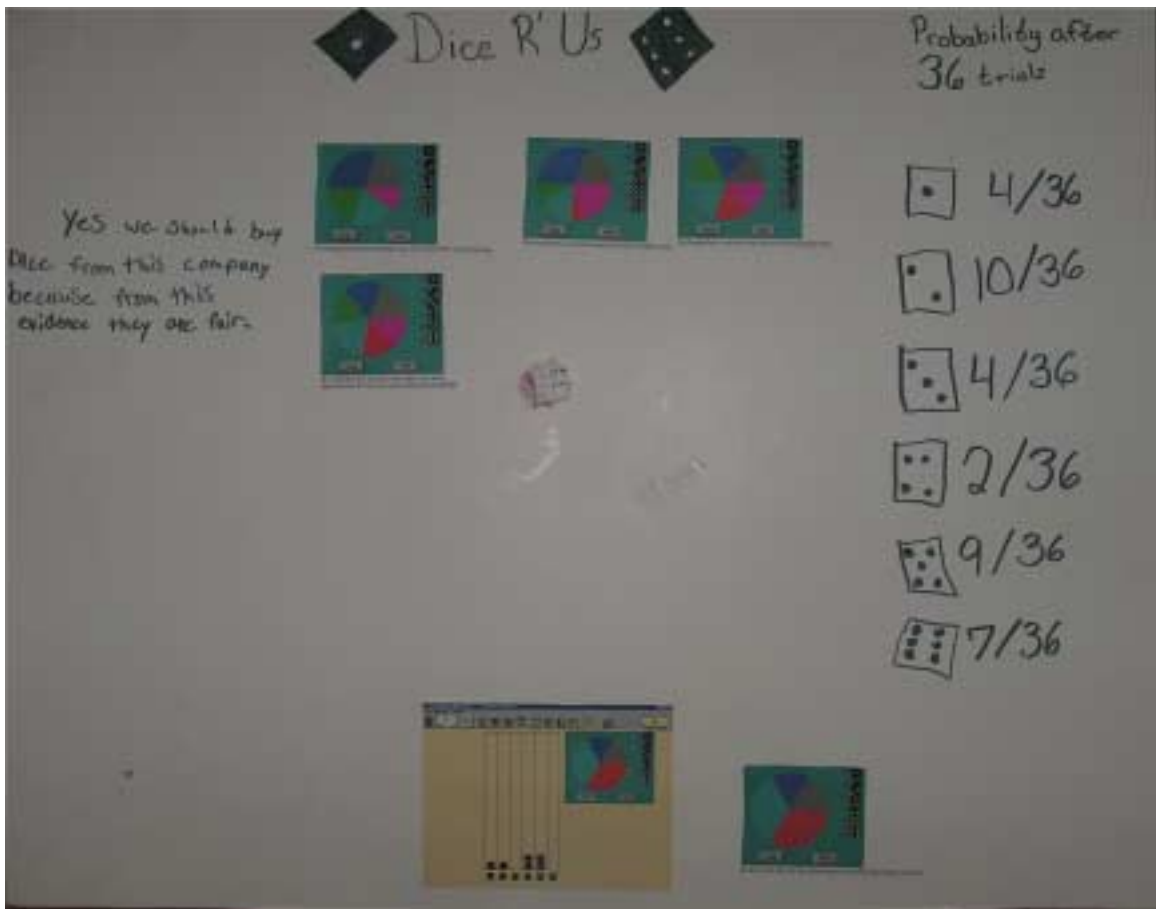


Figure 5. Dannie and Lara's poster.

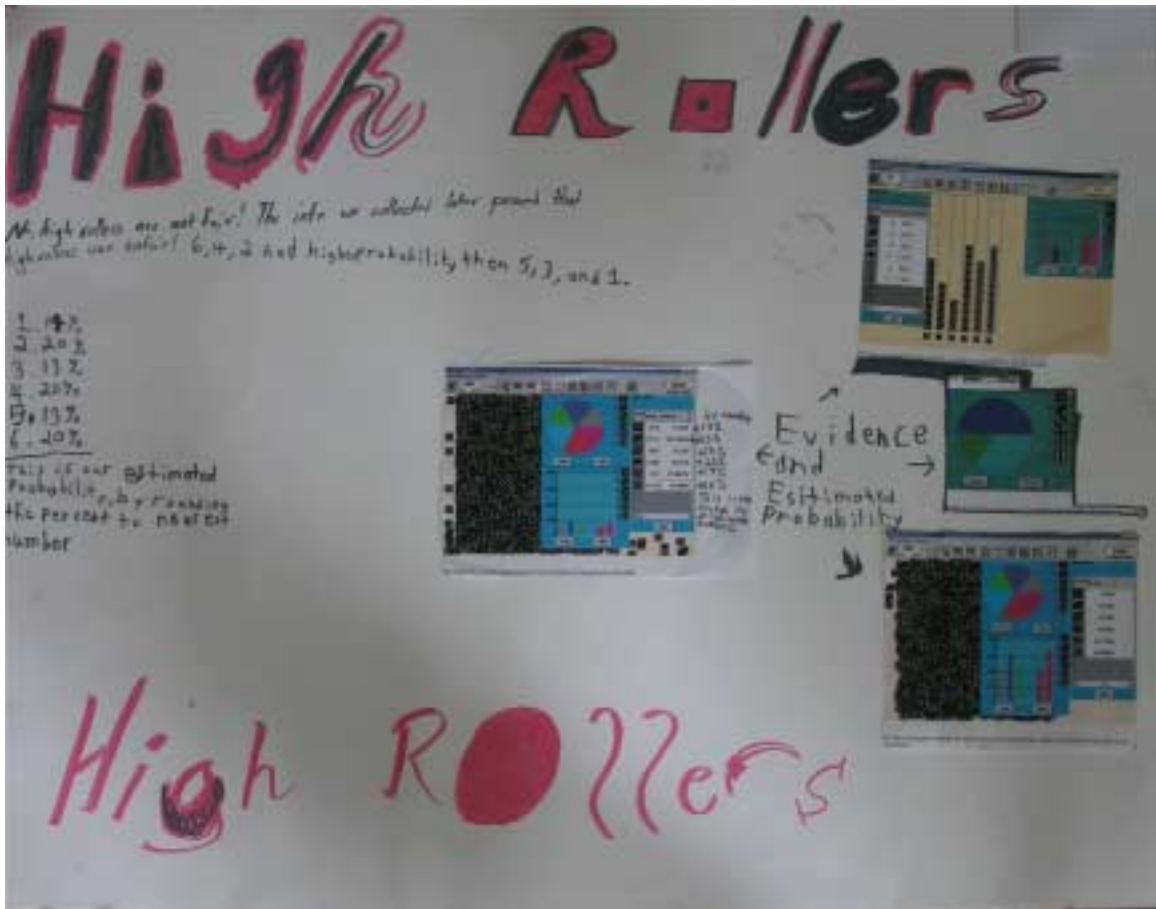


Figure 6. Brandon and Manuel's poster.

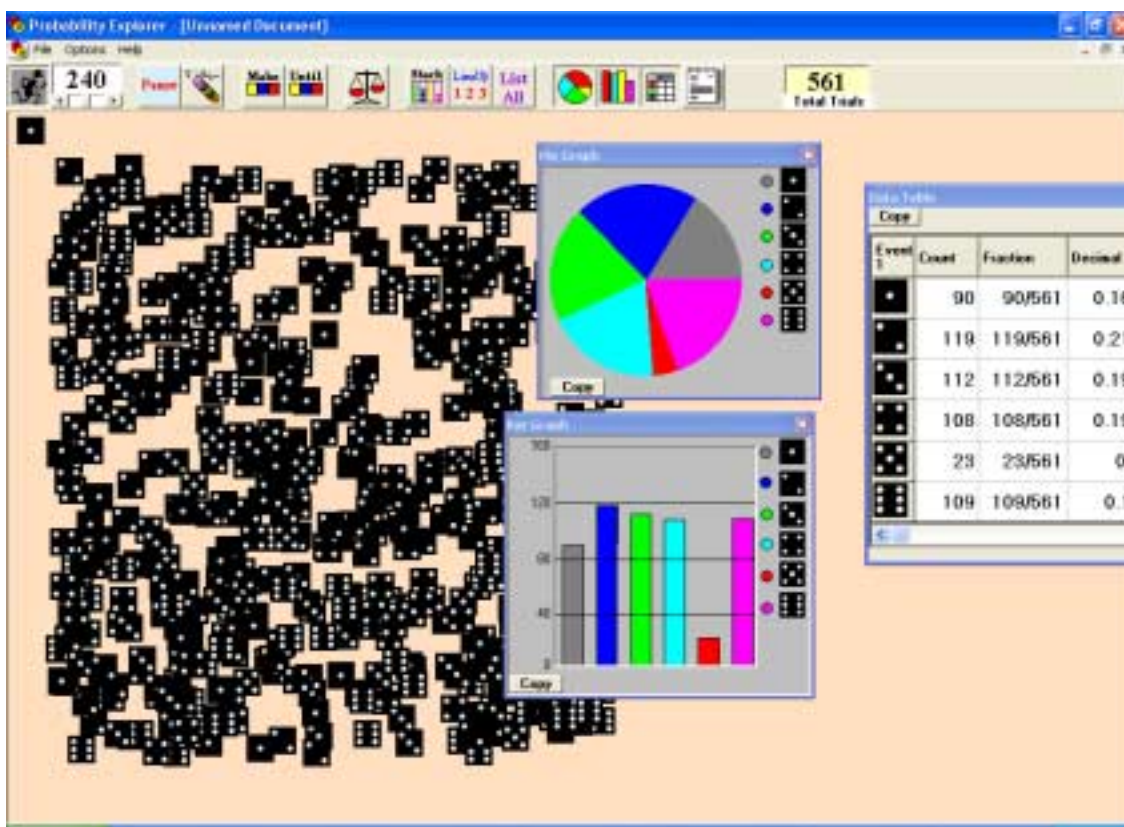


Figure 7. Greg and Jasyn's screen display after 561 trials.

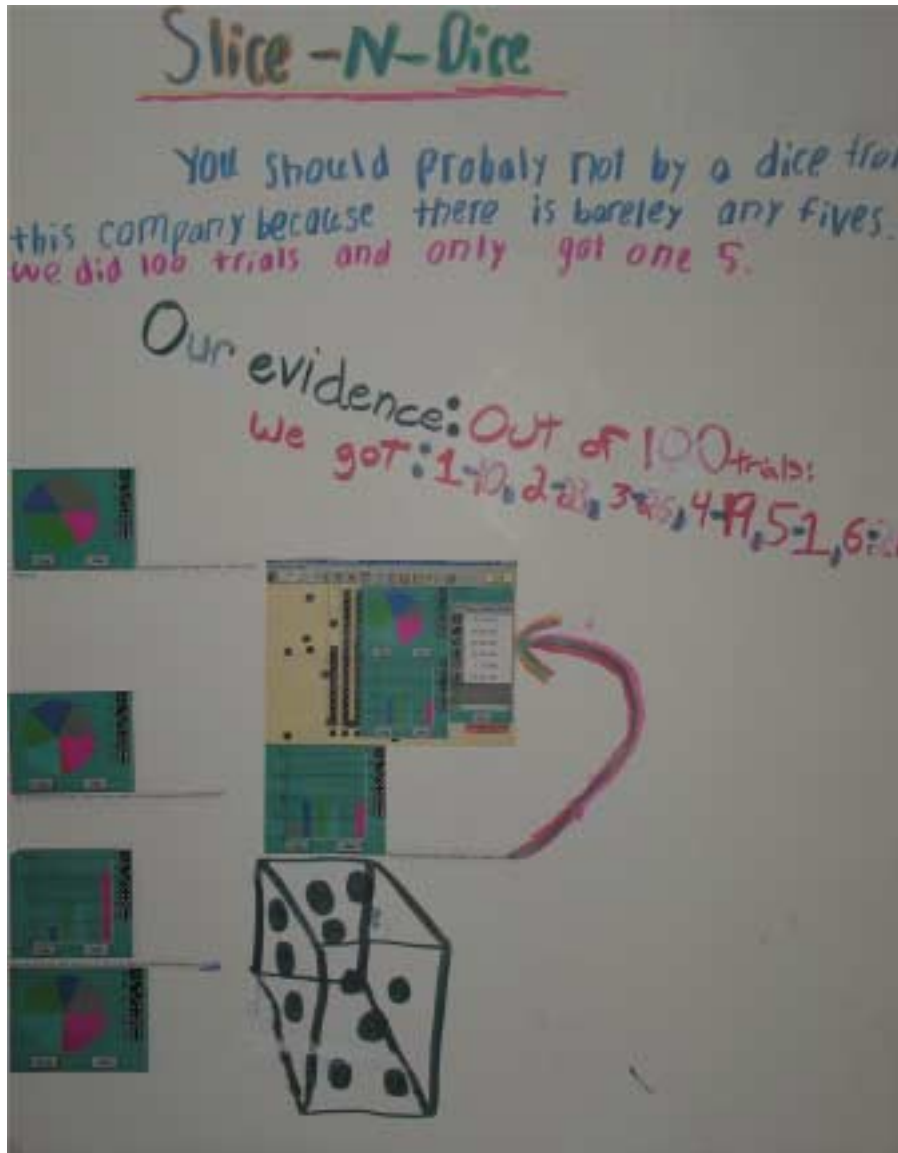


Figure 8. Greg and Jasyn's poster.

Tables

Table 1: Codes Used for Decisions about Data Collection and Display in Each Cycle

Aspect of Cycle	Code
Number of Trials entered in software and What is the cumulative sample size at end of cycle?	Level 1* = 1-40 Level 2* = 41-100 Level 3* = 101-500 Level 4* = 501-1000 Level 5* = 1001+
Is this set of trials (run) independent? Do we add on to previous set of trials or clear and start anew?	I = Independent (clear previous trials) D = Dependent (add on)
What representations do the students choose to have available <i>during</i> the run of trials (i.e., opened either before trial starts or during simulation so that representation changes dynamically)? and What representations do the students choose to have available <i>after</i> the run is complete (i.e., what representations either remain open or are opened and examined in a static mode?)? and What representations do students <i>explicitly refer to</i> verbally and in mouse movements during and after the simulation?	P = Pie Graph FT = Frequency Table RFT = Relative Frequency Table B = Bar Graph S = Stacked Pictogram D indicates during simulation A indicates after simulation
What reasoning regarding variability do students use when making decisions about the data gathered in the cycle? Do the students appear to agree?	WA1 or WA2= Reasoning Within Cycle, Accepting of the Variability, # of students agreeing WU1 or WU2= Reasoning Within Cycle, Unaccepting of the Variability, # of students agreeing AA1 or AA2 = Reasoning Across Cycles, Accepting of the Variability, # of students agreeing AU1 or AU2 = Reasoning Across Cycles, Unaccepting of the Variability, # of students agreeing

Do students save data in word processing application?	S = Save N = Not Save
What inferences do students make about the fairness of the die?	F = Fair (e.g., “that’s fair”) M = Moderately Fair (e.g., “sort of fair”, “not sure yet”) U = Unfair (e.g., “no way”, “we are so unfair!”) N = No inference verbalized

* Sample size levels were determined in consideration of the theoretical probabilities used in the Schoolopoly task² (see Figure 3) and may not be appropriate classifications in other contexts with different theoretical distributions.

Table 2. Frequencies for the type of claim made when two or fewer, or three or more representations are available to students

	No Claim	Claim of Unfairness	Claim of Moderately Fair and Fair
Two or fewer representations	7	6	14 (M) 4 (F) 18 (M&F)
Three or more representations	3	8	1 (M) 4(F) 5 (M& F)