

# MAKING COMPARISONS BETWEEN OBSERVED DATA AND EXPECTED OUTCOMES: STUDENTS' INFORMAL HYPOTHESIS TESTING WITH PROBABILITY SIMULATION TOOLS

HOLLYLYNNE STOHL LEE  
*North Carolina State University*  
*hollylynne@ncsu.edu*

ROBIN L. ANGOTTI  
*University of Washington Bothell*  
*rrider@uwb.edu*

JAMES E. TARR  
*University of Missouri*  
*TarrJ@missouri.edu*

## ABSTRACT

*We examined how middle school students reason about results from a computer-simulated die-tossing experiment, including various representations of data, to support or refute an assumption that the outcomes on a die are equiprobable. We used students' actions with the software and their social interactions to infer their expectations and whether or not they believed their empirical data could be used to refute an assumption of equiprobable outcomes. Comparisons across students illuminate intricacies in their reasoning as they collect and analyze data from the die tosses. Overall, our research contributes to understanding how students can engage in informal hypothesis testing and use data from simulations to make inferences about a probability distribution.*

**Keywords:** *Statistics education research; Middle school; Technology; Informal inference; Data distribution; Variation*

## 1. INTRODUCTION

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 from years of game-playing experiences often support this assumption. Beginning game players may initially develop notions regarding the observed frequencies of any given number that lead to particular generalizations (e.g., "I don't get six a lot so six 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. Furthermore, this type of reasoning about dice can continue to develop as beginners gain more experiences with rolling dice and observing frequencies. At these moments, they may mentally coordinate their observations with their expectations, and a mismatch between the two may yield a perturbation. A person's expectations of outcomes from repeated rolls of a die may be based on initial ideas formulated from earlier experiences or on assumptions made from examining the physical appearance of the die. In either case, students

may be making decisions naturally either to keep their current hypothesis about the likelihood of an event on a die or to alter this hypothesis in light of sufficient contrary evidence.

One cannot precisely determine whether a four will appear when rolling a regular six-sided die. In the presence of this uncertainty, a probability can be estimated based on theoretical analysis or based on empirical data from rolling a die repeatedly. A frequentist (or empirical) approach to probability, grounded in the law of large numbers, has only recently made its way into curricular aims in schools (Jones, 2005; Jones, Langrall, & Mooney, 2007). Teachers are encouraged to use an empirical introduction to probability by allowing students to experience repeated trials of the same event, with concrete materials and/or through computer simulations (e.g., Batanero, Henry, & Parzysz, 2005; National Council of Teachers of Mathematics [NCTM], 2000; Parzysz, 2003). There is general agreement that research on probabilistic reasoning has lacked sufficient study of students' ability to connect observations from empirical data (probability in reality) and expected results based on a theoretical model of probability to inform their judgments and inferences (e.g., Jones; Jones et al.; Parzysz). In addition, many researchers have begun to tackle a difficult question of how students develop notions of informal inference that include examining data sampled (randomly) from finite populations, and data generated from random phenomena that have an underlying probability distribution which may be unknown (e.g., Makar & Rubin, 2009; Pratt, Johnston-Wilder, Ainley, & Mason, 2008; Zieffler, Garfield, delMas, & Reading, 2008). Analyzing an empirical distribution and interpreting data within a larger context are critical components of being able to make judgments, generalizations, or inferences *beyond the data* (Makar & Rubin) about either a theoretical probability distribution or a finite population. The usefulness of engaging students in informal inference tasks that are based on underlying probability distributions, rather than only finite populations, has been encouraged by Pratt et al. and Zieffler et al.

Our research aims to contribute to the literature on understanding informal inference in probability contexts. We examine how middle school students reason about results from a computer-simulated die-tossing experiment to decide between one of two possible hypotheses: (1) the die is fair (i.e., the die has equiprobable outcomes), or (2) die is not fair or biased (i.e., the outcomes on the die are not equiprobable). Thus, we are explicitly interested in how students compare observed empirical data with expected outcomes from a fair die when using technology tools, including various representations of data. Because we are emphasizing data-based reasoning, we designed an instructional task in which students are not given a physical die to examine, but are asked to use data to make a decision, either to support a hypothesis of equiprobable outcomes or to refute this hypothesis and develop an alternative model of the probability distribution. Supporting or refuting a hypothesis involves some level of convincing, whether it is using evidence to convince oneself, a partner, or a third-party (e.g., teacher, researcher, other students). When working with data, the process of convincing requires representation of data, reasoning about data, and communication to others, all of which are encouraged processes in many curriculum documents (e.g., NCTM, 2000). Within this context, we are specifically investigating the following questions:

1. How do students investigate the fairness of a die in a computer-based simulation and use empirical distributions to support or refute a model of a die that assumes equiprobable outcomes?
2. How do students reason about the sample size and variability *within* and *across* samples of data to make sense of distributions of empirical data and a model of equiprobable outcomes?

## 2. RELATED LITERATURE

### 2.1 BELIEFS ABOUT DICE

Several researchers have analyzed students' understanding and beliefs about 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, many students doubted that each outcome is equally probable. However, most of the students who held a belief that die were fair (59% of 108 students), and almost all students who believed die were unfair, did not use a strategy to collect data to confirm or refute their belief. 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 outcomes, (2) for some trials, or (3) compared to another die. The notion that a die could be fair for some outcomes or trials is aligned with an "outcome approach" as described by Konold (1987, 1995). Pratt (2000) also describes students' reasoning about the trial-by-trial variation they notice with dice (and other devices). He found that students characterize randomness such that the next outcome is unpredictable and that no patterns in prior sequences are noticeable (irregular).

Such pervasive beliefs and reasoning about dice are likely to be 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 mathematics curricular materials. Our research is interested in ways in which students' reason from observed empirical data to decide whether they have enough evidence to support a hypothesis that a die has equiprobable outcomes. Thus, students' experiences and beliefs about dice contribute to, but are not a focal point of, our analysis.

### 2.2 REASONING ABOUT VARIATION IN EMPIRICAL RESULTS

Shaughnessy (1997, 2007) called for more attention to the concept of variability in students' learning of probability and statistics, and several studies (e.g., Canada, 2006; Green, 1983; Watson & Kelly, 2004) have shown that students tend to expect little or no variation in samples from an equiprobable distribution, even with a small sample size (e.g., predict 5 ones, 5 twos, 5 threes, 5 fours, 5 fives, 5 sixes for results of 30 fair die tosses). However, Lidster et al. (1995) and Shaughnessy and Ciancetta (2002) found that many students were willing to accept wide variation across several samples and maintain a belief that events are equally likely even with contrary visual and numerical evidence.

Recent research on students' understanding of distributions points to variation as a key concept and suggests that students often struggle with coordinating variation within and across distributions when reasoning about center of a distribution (e.g., Reading & Reid, 2006). Shaughnessy (2007) delineated eight aspects of variation that arise in various contexts, including variation over time, variation over an entire range, and variation between or among a set of distributions. Mooney (2002) also described ways of reasoning about variation that include (1) comparing frequencies of outcomes *within* a data set to each other, (2) comparison to an expected theoretical distribution, (3) comparison to an expected distribution based on intuition of the context, and (4) comparing empirical distributions *across* independent data sets to examine *differences* in variation (especially with small numbers of trials). Watson (2009) examined students' developing understandings of variation and expectation for grade 3 through 9 students. Her results suggest that students may be able to appreciate and anticipate variation in data well before they develop a natural tendency to anticipate expected distributions or predicted values

based on a theoretical distribution. However, many students in grades 7 and 9 were able to integrate ideas of expectation and variation.

There has been little consideration in prior studies of students' attention to *invariance across distributions* when comparing the variation present in each distribution. This type of reasoning across data sets is important if students are trying to use data from repeated samples to make inferences about an unknown theoretical probability distribution from a frequentist perspective. This type of reasoning is also valuable when trying to reject a hypothesis of a particular probability distribution (e.g., whether or not a coin or a die has equiprobable outcomes). Pratt and colleagues (2008) posed tasks to 10–11 year-olds similar to the one used in our study. Students used a simulation tool to collect data about a weighted die and were asked to describe the die based on examining data, where they did not have access to the hidden probability distribution. Pratt and his colleagues found that students seemed to be looking for invariance across samples of data but were often frustrated because they tended to focus more on what was changing and did not appear to instinctively know or come to realize the role of sample size.

### 2.3 USE OF SIMULATIONS AND REPRESENTATIONS

Although many organizations such as NCTM (2000) encourage the use of technology-based simulations, research on the role of simulation software in teaching and learning probability is a relatively recent endeavor (Jones et al., 2007). Using simulation software affords students an opportunity to observe the dynamic accumulation of data in numerical and graphical forms while data are being generated. This visualization of data can be a powerful motivator for students noticing variability in short and long term behavior of random events.

Representations available in computing environments can afford or hinder students' work in examining empirical data from simulations. For many years, researchers and teachers have examined how students' use of multiple representations can help them understand mathematical concepts. In particular, Ainsworth (1999) suggests that representations can be used in complementary ways to provide the same information in a different form and to provide different information about the same concept. Many researchers studying the use of simulation tools in probability learning have incorporated different representational forms in the software and have examined how students used these representations. For example, in early research by the authors (Drier, 2000a, 2000b; Stohl & Tarr, 2002), students used representations (e.g., bar graph, pie graph, data table) in a computer 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 (law of large numbers). Students often recognized that larger numbers of trials resulted in distributions that more closely resembled their expectation from theoretical probabilities used to design an experiment. Others have documented how students are able to make connections between distributions of data from a simulation and a theoretical distribution, with many students, though not all, attending to the effect of the number of trials (e.g., Abrahamson & Wilensky, 2007; Ireland & Watson, 2009; Konold & Kazak, 2008; Pratt, 2000, 2005; Pratt et al., 2008; Prodromou, 2007; Prodromou & Pratt, 2006). Collectively, these studies suggest that these tools give students control over designing experiments, conducting as many trials as they desire, viewing and coordinating various representations, and helping students develop understandings between expectations of outcomes based on a theoretical model and observations of empirical data, and how observations of empirical data may inform development of a theoretical model.

### 3. ANALYTICAL FRAMEWORK

The focus of our research is on how students collect and analyze empirical data and compare their results with expectations based on a model of the probability distribution of a die that assumes equiprobable outcomes. In essence, the task is designed to engage students in the process of a statistical investigation to informally test a hypothesis of equiprobable outcomes and make an informal inference about an underlying probability distribution. The students are not engaging in formal hypothesis testing techniques, however, the task is aligned with those suggested by Zieffler and colleagues (2008) that can engage students in informal inference by eliciting a judgment about which of two competing models or statements is more likely to be true.

There are several well-developed frameworks describing the process of statistical investigation; for example, the PPDAC model used by Pfannkuch and Wild (2004) or the PCAI model used by Friel, O’Conner, and Mamer (2006). These authors pay attention to the ideas of collecting data through samples (particularly random samples) and the importance of distribution and variation in the analysis of data. The components of a statistical investigation may emerge linearly, though not necessarily, and may include revisiting and making connections among the components. The first component includes *Posing a question*. Such questions should be about particular contexts and should be motivated by describing, summarizing, comparing, and generalizing data within a context. Most questions are in the form of a conjecture or hypothesis about a context. *Collecting the data*, the second component, includes a broad range of collection opportunities from populations and samples, including those generated from a probability experiment or simulation. The third component of the model, *Analyzing data*, should encompass perceiving a set of data as a distribution when a probability experiment or simulation is conducted. This component also includes describing and analyzing variation, as well as organizing and displaying data in charts, tables, graphs, and other representations. *Interpreting the results*, the fourth component, involves making decisions about the question posed within the context of the problem based on empirical results in relation to sample size. This includes making judgments and inferences about a probability distribution.

To answer our research questions, our analysis is focused on the students’ work while engaged in informal hypothesis testing through the collecting, analyzing, and interpreting components of their investigation that occur in a cyclic manner. Thus, we need to explicate a framework that will focus our analysis (see Figure 1).

In an informal hypothesis testing context, an initial hypothesis of a distribution ( $H_0$ ) is generated from the problem context (e.g., the die is fair—the distribution is uniform) in the *Pose Question* phase. In the next phase, *Formulate Expectations*, the initial assumptions of a probability distribution provide an image of what students expect to observe in empirical data under the hypothesis (e.g., if the die is fair, each outcome should occur about the same number of times). Students then *Collect Data* to test the hypothesis. Depending on the context of the task, students may be guided in how to collect the data, or students may be asked to make all the data collection decisions. In the two phases labeled *Analyze* (Figure 1), students would ideally be analyzing the data collected and then comparing the results against their expectations and initial hypothesis about the probability distribution (Watson & Kelly, 2004). Noticing patterns in the data may prompt them to make *Interpretive Decisions* such that they may question the prior assumptions, or they may not believe the data vary enough from their expectations to contradict their initial assumption. Their reasoning may then lead them to decide to keep  $H_0$  or to form an alternative hypothesis ( $H_A$ ). In either case, they could complete the investigative cycle and make a final *Interpretation* or continue in the cycle to *Formulate Expectations*, *Collect Data*, and *Analyze* to again test the reasonableness of the match between their expected results based on a hypothesized probability distribution and an observed distribution from repeated empirical trials. Of course students may also decide to collect more data because it is pleasurable or playful (as has been observed in Drier, 2000a; Lee, 2005). However, even these playful excursions provide

opportunities for students to observe additional data distributions. Whether their cycles of data collection are purposeful or playful, at some point students need to make a final claim about either supporting or refuting the original hypothesized distribution, and possibly to make inferences about an alternative distribution. In a dynamic simulation environment where data are displayed in various representations while being generated, students may engage in analyzing and interpreting *while* data are collected (Drier, 2000b; Stohl & Tarr, 2002; Pratt 2000, 2005; Pratt et al., 2008). Thus, the components of a statistical investigation can become integrated through several iterations of data collection, as shown in Figure 1.

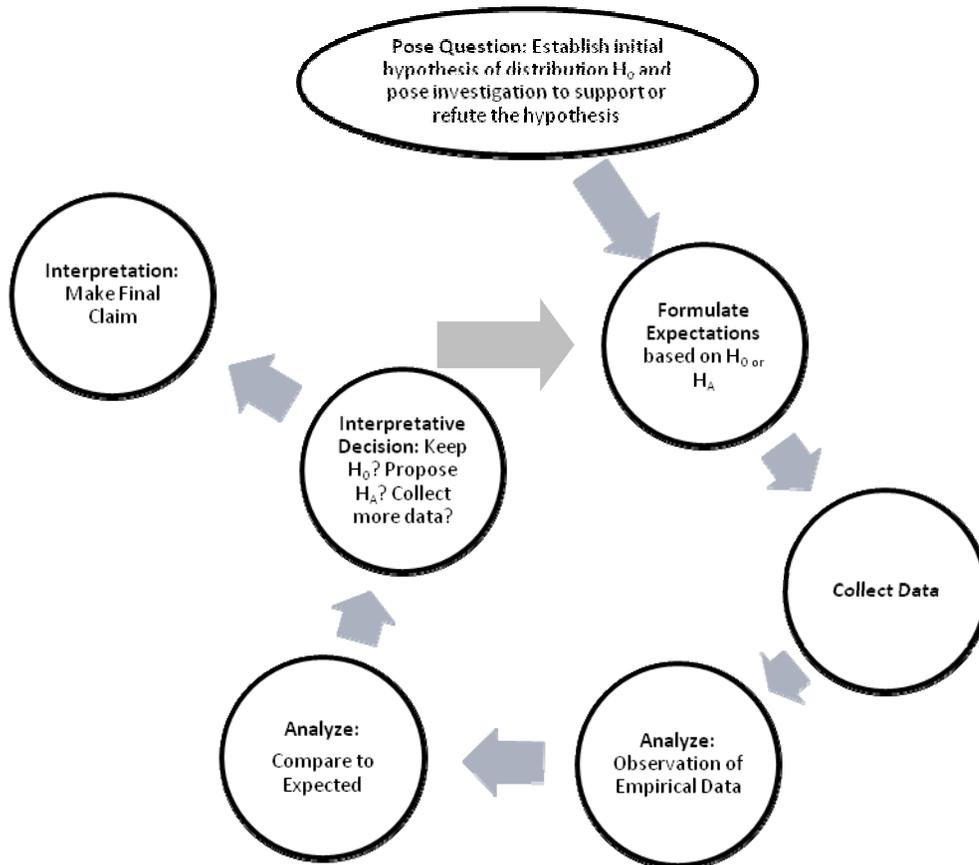


Figure 1. Informal hypothesis testing cycle in a statistical investigation that focuses on comparing expectations and observations

The robustness of students' analysis of observed empirical data with respect to their expectations based on initial assumptions of a probability distribution may be influenced by their understanding of the independence of trials, variability of their data, and sample size. Students need to consider that different trials and different sets of trials (samples) are independent of one another and variability among individual trials and samples is to be expected. They also need to coordinate conceptions of independence and variability with the role of sample size in the design of data collection and interpretation of results. Relative frequencies from larger samples are likely to be more representative of a probability distribution whereas smaller samples may offer more variability and be less representative. For example, ten rolls may yield no 3's and such data may support a child's notion that rolling a 3 is an improbable (or even impossible) event, although such a claim would likely not be made by someone who had made a more robust connection

between observed data and expectations based on a theoretical probability distribution, and the importance of sample size.

Pratt et al. (2008) called for a need to “conjecture about what the children might be *expecting* to see” (p. 124, italics in original). Students’ hypothesis about a probability model and their subsequent expectations from experimentation may or may not be stated explicitly. To help account for this in our analysis, we employ the constructs of *external* and *internal resources* used in Pratt’s (2000) study on students’ probabilistic reasoning. Internal resources consist of an individual’s understandings, including intuitions and beliefs. External resources are those “that reside outside of the individual but within a setting” (p. 605), and would include the social interactions, available tools (including various representations available in a software tool), and the context of the mathematical task. 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 cannot directly observe a learner’s internal resource. However, by analyzing the interactions with external resources, we as observers can make inferences about a learner’s internal system. Specific to our research, we are using students’ actions with the software and their social interactions to infer their expectations of results based on a hypothesis of likelihood of events (e.g., equiprobable outcomes) and whether or not they are formulating a new hypothesis.

## 4. METHODOLOGY

### 4.1 CONTEXT OF STUDY

The study took place during the first month of sixth grade (age 11-12) in an average-level mathematics class ( $n = 23$ ) in an urban, public middle school in the southern United States. The class was engaged in a 12-day probability unit designed and taught by Lee and Tarr. During class, students worked in groups of 2-3, and used real objects (coins, dice, and spinners) and *Probability Explorer* (Stohl, 2002) on laptops.

The software enables users to specify the number of trials, update displays of results after each trial, and view simulation data in iconic images on the screen which can be stacked in a pictograph or lined up in the order in which they occur. Additionally, data can be viewed in a pie graph (relative frequency), bar graph (frequency), and data table (frequency and relative frequency). The multiple representations serve complementary functions (Ainsworth, 1999) in which some representations differ in the information that is expressed (e.g., frequencies in a bar graph versus the relative frequency table) or show the same information in different ways (e.g., frequencies in iconic pictogram, bar graph, and frequency table).

### 4.2 INSTRUCTIONAL SEQUENCE

The six tasks used in the 12-lesson unit required students to carry out computer-based simulations, and to collect, display, and analyze data in order to draw inferences and formulate convincing arguments based on data. Each lesson was about 60 minutes in length. Whole class discussions were designed to elicit students’ reasoning about observed data collected through empirical trials and expectations based on a model of the probability distributions. Throughout the 12 days we encouraged dispositions towards using an investigative cycle and norms about making data-based arguments, mainly to their peers, but also to the teachers/researchers when occasionally prompted during small group work or during whole group discussions.

The first two tasks (Fair Coin Tosses and Fair Die Tosses) were designed to help students make connections between random events with familiar objects to the random events generated in the software. These tasks purposely posed questions to elicit students' primary intuitions (Fischbein, 1975) about equally-likely outcomes, randomness, and theoretical probability so we could build upon those intuitions throughout the unit. In both tasks, students had extensive experience in using the software to simulate fair coin tosses and die rolls, examining many samples of small and large trials with the various representations (table, stacking columns, bar graph, pie graph), and discussing variability within and across samples as a whole class. The third task provided students with a situation where some elements of the sample space were known to students (Mystery Marble Bag with  $N$  marbles) and asked students to collect data and draw inferences about the theoretical distribution.

The fourth task (Mystery Fish in a Lake) was designed to induce a perturbation about how to determine the number of trials to run when the population of fish was unknown. The task also served the purpose of encouraging students' social negotiations about whether they were confident that enough data had been collected to estimate the probability distribution. In the fifth task, students used the Weight Tool to enter estimates for the probability of each event on a physical spinner and used empirical data to test the "goodness" of their estimates. These last two tasks aimed to draw students' attention to the role of sample size in comparing empirical data with a theoretical model of probability.

The Schoolopoly task (Figure 2) — the focus of our analysis in this paper — spanned the last three days of instruction. The Schoolopoly task was included in the framework offered by Zieffler et al. (2008) as an exemplar of a task that can elicit students' thinking about informal inference. In this task, students collected evidence to infer whether or not a dice company produces dice with equiprobable outcomes. For each dice company the teachers designed the computer simulation by assigning probabilities to each of the six outcomes in the Weight Tool (see Table 1). The Weight Tool was then hidden from students during the activity, thus creating a context in which there exists an unknown probability distribution, similar to the task used in Pratt et al.'s 2008 study. In this task, students were asked to consider a rumor that some of the dice may be biased, to justify whether they found enough evidence to retain an equiprobable hypothesis or to claim that the dice had non-equiprobable outcomes, and to propose an estimate for the probability distribution (see Questions 2 and 3 in Figure 2). The students had to decide how much data to collect, how to arrange the iconic form of the outcomes on the screen, how to display data numerically and graphically during analysis, and what data sets and representations to use to make and support claims about the whether the dice were fair with equiprobable outcomes. Because of students' prior experiences with fair dice and the challenge posed in the Schoolopoly task, we assumed students had a reasonable understanding of equiprobable outcomes and could form reasonable expectations of an empirical distribution.

In order to promote students' perception of the authenticity of the Schoolopoly task, we purposely had six different dice companies vying for the "bid" with each company producing dice with different biases and only one having equiprobable outcomes. The task was designed to elicit classroom conversations about which of the six dice companies to recommend for use in the game and encouraged whole-class consensus building after each group had prepared the poster report to the school board. We discuss some of the beneficial aspects of the Schoolopoly task and a broader perspective on classroom results in a different paper (Tarr, Lee, & Rider, 2006).

### Schoolopoly

Your school is planning to create a board game modeled on the classic game of *Monopoly*<sup>TM</sup>. The game is to be called *Schoolopoly* and, like *Monopoly*<sup>TM</sup>, 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*

*Table 1. Weights and theoretical probabilities (unknown to students) for events 1-6 in each company*

Company Name	Weight [P(1)]	Weight [P(2)]	Weight [P(3)]	Weight [P(4)]	Weight [P(5)]	Weight [P(6)]
Dice R’ Us	2 [0.125]	3 [0.1875]	3 [0.1875]	3 [0.1875]	3 [0.1875]	2 [0.125]
High Rollers, Inc.	2 [0.133]	3 [0.2]	2 [0.133]	3 [0.2]	2 [0.133]	3 [0.2]
Slice n’ Dice	4 [0.16]	5 [0.2]	5 [0.2]	5 [0.2]	1 [0.04]	5 [0.2]

### 4.3 THREE FOCUS PAIRS

Prior to instruction, three focus pairs of students were selected based on scores on a standardized mathematics achievement test as well as a pretest on probability concepts developed by the researchers (e.g., determining a priori probabilities, comparing probabilities of two events, using data to make statements regarding probability). With both pretests, we grouped the scores in thirds (high, middle, low). Candidate focus students were chosen by their relatively consistent ranking on both pretests. Six students were chosen to be collectively representative of gender and ethnicity of the class.

Dannie and Lara (Caucasian girls) represented the high-scoring group who investigated the Dice R' Us company. 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 Table 1 to examine the difficulty levels of the theoretical probability distributions for each company. The slightly biased die from the Dice R' Us company was given to Dannie and Lara in anticipation that they had developed an appreciation for sample size [based on their work on previous tasks] and would develop the realization of the need to collect a large number of trials to detect bias in the die.

For the purposes of our analysis with three pairs of students, each testing a different virtual die, we recognized that the overall instructional decision for using differently-biased dice with each pair added a layer of complexity to our analysis for comparing across groups. Thus, our comparisons were not about the quality of the “correctness” of the students’ decision to confirm or refute that the die has equiprobable outcomes or their ability to estimate the probabilities for each outcome based on their data collected. Rather, we focused on *how* students used the software to conduct their investigation, how they compared observed empirical data with expected results, and the reasoning they communicated with each other to support their claims.

#### 4.4 SOURCES OF DATA

All students in the class were seated in pairs or groups of three at tables with a PC laptop, calculators, and manipulative materials (e.g., dice, spinners) readily available. The three focus pairs worked at tables where their laptop computers were connected to a PC-to-TV converter to video-record their computer interactions while microphones captured their conversations. Detailed video and audio recordings were only made of these three pairs. In addition, there was a wide-angled video camera focused on the three tables to capture students’ social interactions with each other and the teachers-researchers and two additional cameras capturing the whole class. 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, students were routinely saving data and making notes during data collection in a word processing document. Each pair also constructed a poster with answers to questions in Figure 2 and appropriate data displays as evidence to support their claims. Thus, data considered for analysis included video with annotated transcripts, students’ word processing documents, and students’ posters.

#### 4.5 ENACTING OUR FRAMEWORK FOR ANALYSIS

Using our understanding of results from related studies, Pratt’s (2000) notion of internal and external resources, and our guiding framework in Figure 1, we analyzed how pairs of students coordinated the empirical data with a hypothesized probability distribution while conducting their investigation. An initial analysis of the videos and annotated transcriptions identified critical events (Powell, Francisco, & Maher, 2003) in the students’ work. Our initial analysis was generative in nature and brought students’ key decisions and actions in data collection and analysis to the fore (Strauss & Corbin, 1990). Five major decisions/actions emerged as salient across the three pairs’ work. Thus, in order to analyze students’ work more closely, we broke each pair’s work into investigation “cycles” similar to those depicted in Figure 1 that included their decisions and actions to: 1) Carry out a number of trials, 2) view representations (pie, bar graph, data table, etc) during and after the simulation, 3) save the data (or not) as evidence for further analysis and justification, 4) make a conjecture based on data to reject (or not) the initial hypothesis of equiprobable outcomes, and 5) proceed to collect more data. In addition, we noted that when comparing the empirical distributions to their expectations, the students reasoned about variability in the empirical distributions in two major ways, within a data set, and across data sets.

When reasoning within a data set, they seemed to either accept that the variability was not different enough from their expectations, or that it was different enough. When reasoning across several empirical distributions, they were either attending to variance (differences) or invariance (similarities).

Each of the transcription documents was segmented into distinct cycles and coded (Miles & Huberman, 1994) for students' actions/decisions in the five major categories and their reasoning about variability. For example, for each cycle we coded which representations (S=Stacked pictograph, B=Bar Graph, P=Pie Graph, FT=Frequency Table, RFT=Relative Frequency Table) students utilized to view the data, either during (D) or after (A) data collection. We coded whether they made an explicit interpretive decision about whether to refute (R) or confirm (C) the equiprobable assumption, or if they were unsure (U). In developing meaningful codes for sample size, we considered the theoretical probability distributions for the three companies and chose sample sizes that reflected reasonable margins of error and would allow increased confidence in predictions as the sample size levels increased. The five sample size levels used for all pairs were 1) 1-40, 2) 41-100, 3) 101-500, 4) 501-1000, 5) 1000+. The lowest sample size level has a maximum of 40 to account for even the most biased dice company's (Slice n' Dice) distribution. See the Appendix for detailed codes.

The coding process resulted in a condensed description of each pair's work that could enable us to quickly summarize each cycle (see Appendix). This summary allowed us to focus more clearly on critical events, that is, those events that were relevant for understanding how students were comparing the empirical data to their expectations based on a hypothesized theoretical model of probability (Figure 1). We then tallied the occurrences of each code (Table 2) to examine patterns within and across pairs.

## 5. RESULTS

Recall our research questions: "How do students investigate the fairness of a die in a computer-based simulation and use empirical distributions to support or refute a model of a die that assumes equiprobable outcomes?" and "How do students reason about the sample size and variability *within* and *across* samples of data to make sense of distributions of empirical data and a model of equiprobable outcomes?" We discuss our findings in two parts. First, we provide an overall picture of the pairs' investigation based on their use of representations, use of sample size, reasoning about variability, and the conjectures that they made during their investigation. Secondly, we provide a closer analysis in which we present how students working together compared empirical data and expected results as they went through the cycles (see Figure 1) in their investigation, focusing on those events that seemed most critical from our coding of cycles. It is within this closer analysis of the pairs' work that we are considering the individuals' reasoning and their mental images of expected results as inferred by the researchers from their use of the external resources and social interactions with each other.

### 5.1 SUMMARY OF CASES

Students' work on this task spanned three instructional days. The first day consisted of posing the task and pairs working on collecting data, recording results and thoughts in a word processing document, including copying and pasting selected representations from the software. The second day included some additional data collection as well as the creation of posters. The third day included finishing posters and presentations by each group. Our analysis in this paper does not include the presentations.

Table 2 contains a summary of the type of decisions and actions from each pair's work including their reasoning about variability within and across samples, and conjectures they made.

Table 2. Summary of Focus Pairs' Work on Schoolopoly Task

		Dannie & Lara Dice R Us	Brandon & Manuel High Rollers	Greg & Jasyn Slice n' Dice
Number of Cycles Completed		20	8	11
Cumulative	L1: 1-40	11 (55%)	0 (0%)	6 (54.5%)
Sample Size in Each Cycle	L2: 41-100	3 (15%)	4 (50%)	4 (36.4%)
	L3: 101-500	3 (15%)	2 (25%)	0 (0%)
	L4: 501-1000	2 (10%)	0 (0%)	1 (9.1%)
	L5: 1001+	0 (0%)	2 (25%)	0 (0%)
Representations Used <i>During</i> Data Collection	0 Representations	3 (15%)	0 (0%)	0 (0%)
	1 Representations	9 (45%)	1 (12.5%)	2 (18.2%)
	2 Representations	8 (40%)	1 (12.5%)	1 (9.1%)
	3 Representations	0 (0%)	2 (25%)	8 (72.7%)
Representations Used <i>After</i> Data Collection	4 Representations	0 (0%)	4 (50%)	0 (0%)
	0 Representations	0 (0%)	0 (0%)	2 (18.2%)
	1 Representations	8 (40%)	1 (12.5%)	5 (45.4%)
	2 Representations	12 (60%)	2 (25%)	2 (18.2%)
Conjecture Made of Whether the Die is "Fair" <sup>a</sup>	3 Representations	0 (0%)	2 (25%)	2 (18.2%)
	4 Representations	0 (0%)	3 (37.5%)	0 (0%)
	No conjecture	6 (30%) <sup>a</sup>	0 (0%)	3 (27.3%)
	Confirming hypothesis (i.e., Fair)	4 (20%) <sup>a</sup>	4 (50%) <sup>a</sup>	0 (0%)
Reasoning About Variability <sup>a</sup>	Unsure	9 (45%) <sup>a</sup>	2 (25%) <sup>a</sup>	0 (0%)
	Refuting hypothesis (i.e., unfair)	3 (15%) <sup>a</sup>	3 (37.5%) <sup>a</sup>	8 (72.7%)
	Not Explicitly Stated	9 (45%) <sup>a</sup>	1 (12.5%) <sup>a</sup>	6 (54.5%) <sup>a</sup>
	Within Samples (Comparing frequencies)	Tolerant of differences	7 (35%) <sup>a</sup>	4 (50%) <sup>a</sup>
Not tolerant of differences		4 (20%) <sup>a</sup>	6 (75%) <sup>a</sup>	5 (45.4%) <sup>a</sup>
Across Samples (Comparing distributions)	Attention to variation	1 (5%) <sup>a</sup>	2 (25%) <sup>a</sup>	0 (0%)
	Attention to invariance	0 (0%)	0 (0%)	2 (18.2%) <sup>a</sup>
Final Interpretation and Claim on Poster	Refute or Confirm Initial Hypothesis (Is die fair?)	Confirm	Refute	Refute
	Approximate Distribution matches Claim?	No—gave non-uniform distribution from 36 trials	Yes	Yes—but gave frequency distribution from 100 trials

<sup>a</sup>These numbers do not sum to the total (or 100%) for each pair because more than one conjecture or statement regarding variability was sometimes made by one or both students during a single cycle.

Although each pair worked the same amount of time, there were differences in the number of cycles of data collection each group employed, ranging from 8 to 20 cycles. Aspects that

contributed to the number of cycles were students' choice of sample size (smaller samples can be collected faster), use of basic technology of cutting and pasting to save data as evidence to support their conjectures, and amount of time students spent discussing collected samples. Overall, two of the three pairs' cycles of work resulted in cumulative sample sizes that were 100 or below (Levels 1 and 2). However, as elaborated in the pair analysis, the instances where students employed a large sample size had substantial impact on their reasoning about the task. Further, Brandon and Manuel and Greg and Jasyn had a relatively high percentage of cycles in which they utilized three or more representations either during or after data collection.

Due to the description of the Schoolopoly task and work done prior in the instructional unit with fair dice, we inferred that the students all had an image that the probability distribution for the outcomes of an unbiased die should be equiprobable (i.e., uniform). Thus they were able to initially approach the problem with an image of equiprobable outcomes and were asked to collect and analyze empirical data to confirm or refute that the die had equiprobable outcomes (hereafter EPO).

Dannie and Lara and Brandon and Manuel (the pairs with the less obviously biased dice) each had at least one cycle when more than one conjecture code (C, U, R; see Table 2 and Appendix) was recorded. This occurred when students disagreed and expressed conflicting conjectures, or when the students made conjectures at more than one point during a cycle (e.g., claiming "sort of fair" after 40 trials but by 200 trials claiming the die was "fair"). There were also nine cycles when no student made any verbal conjecture about whether or not to refute the EPO assumption.

## 5.2 PAIR ANALYSIS

In this section, we delve deeper into the ways in which the pairs worked through the Schoolopoly task. Using the framework in Figure 1, we are able to describe better some of the details of their work and make conjectures about ways they are reasoning about the external resources that may allow us to hypothesize their internal resources or images.

**Pair 1: Dannie and Lara (*Dice R' Us*)** Most of Dannie and Lara's 20 cycles of data collection and analysis were similar, using Level 1 sample sizes (see Table 2), making the same conclusion of "sort of fair," utilizing the pie graph both during and after data collection, and reasoning about variability occurring *within* a cycle. Twenty-five percent of their cycles contained cumulative sample sizes greater than 100 trials. Aside from the pie graph, the other representations used were the stacked columns (35% of cycles), the frequency table (20% of cycles), with the relative frequency table only being utilized on the last cycle. Even though they used four different representations, they never used more than two representations during a cycle, thus limiting their view of complementary representations of a data set (Ainsworth, 1999).

During the first few cycles, Dannie expressed interest in only saving data sets that were relatively "even." From that, we can infer that she had an initial image of the equally distributed outcomes of the die and that she was anticipating the resulting data to support this image. Lara, on the other hand, thought they should make decisions to save data "if it looks kind of even or if it looks really not even." Thus, early on, Dannie appeared to have a goal to confirm the EPO assumption whereas Lara had a goal to examine data for evidence of whether or not to refute this assumption.

In the first four cycles, they collected 10 trials in each cycle, adding on to the previous cycle, resulting in a cumulative sample size of 40 in Cycle 4. They both made tentative conjectures that they were unsure ("sort of fair" and "about even") whether the data represented an equiprobable distribution. In the first 20 trials there was a low number of fives and this appeared to make the students hesitant to make a definitive judgment. They were reasoning within each cycle of data and making comments related to the variability between frequencies of each outcome particularly focusing on the low number of fives.

Dannie expressed an initial image of expected results based on EPO that seemed to influence her tolerance of the variation visible in the representations, often resulting in her arguing that the data appeared “even.” Although Lara also initially had an expectation based on EPO, she was more willing to compare the data against her expectations to question the feasibility of equiprobable outcomes. Thus, she was not wholly convinced by Dannie’s argument to accept the label of “even” to describe the data distribution in the first six cycles. However, in Cycle 7 (100 trials), with data displayed in a pie graph which visually appeared to have equal slices, the partners agreed that the data appeared “really even.” Cycle 7 was the only instance of Lara comparing across distributions, between this data set and the earlier one that had a low number of fives. This instance seemed to reinforce Dannie’s image of equiprobable outcomes and allowed Lara to accept that these empirical data did not provide evidence for refuting the EPO hypothesis. The students did not make any explicit reference to the influence of sample size (100 trials) on their reasoning.

Lara was the first to explicitly refer to sample size when she disagreed with Dannie’s conjecture of refuting the EPO assumption (“it isn’t fair”) in Cycle 13. Lara contended that the conjecture could not be made because the sample size was only 12 (“we don’t know if it’s fair, because that was 12 [trials]”). Interestingly, this was the first instance of Dannie’s willingness to consider that her image of the EPO did not match the empirical data. However, she seemed perturbed that this mismatch could exist and in subsequent cycles she began a quest “to see how long it takes before the die is fair,” perhaps because of a strong belief that die should be fair (Green, 1983; Watson & Moritz, 2003). Dannie was now evoking the significance of sample size to support her contention that the distribution of empirical data would approach the theoretical distribution which she expected to be equally likely. Similar to results of previous studies (e.g., Lidster et al., 1995; Shaughnessy & Ciancetta, 2002), this student did not seem to consider that a distribution of data could provide evidence to refute the EPO hypothesis.

The partners continued to collect more evidence and coordinate their images with the empirical data. Dannie, being quite tolerant of any variability and continuing to conduct more trials, only saved the results that support her EPO image. A disagreement between the partners occurred when the sample size exceeded 500 (Level 4). Dannie explained to the teacher/researcher that their strategy was to see “how long it takes until it gets fair” and admitted that, after 834 trials, she did not “yet” consider it to be fair. This further confirmed her expectation that the distribution of the empirical data would eventually match her EPO image. However, after the sample reached 906 trials with data shown in the pie graph and the frequency table, they decided to save both representations and make a conjecture. Lara conjectured to refute the assumption of EPO (“after 906 trials it wasn’t fair”) whereas Dannie still negotiated for “pretty fair” or “sort of fair.” Dannie pointed out the distinct variations among the frequency of each outcome in the table (“these three, the three, four, and five, they have more rolls than the one, two, and six”) but this variation was not great enough for her to acknowledge that the pattern in the data contradicted an assumption of EPO. Her willingness to tolerate this variation was again similar to findings from Lidster et al. (1995) and Shaughnessy and Ciancetta (2002).

The external resources of the pie graph and frequency table, along with the large sample size (906 trials), seemed to support Lara’s reasoning towards a conjecture about refuting the EPO hypothesis. Lara was coordinating sample size and the stabilization of certain outcomes occurring less than others (variation within sample); however, her social interactions with Dannie resulted in Lara’s conjecture of refuting the EPO assumption being overshadowed by Dannie’s decision to report the die was “fair” on their poster.

While creating their poster (Figure 3), Dannie and Lara indicated they had not estimated the probabilities of the outcomes of the die (see Question 3 in Figure 2). Instead of reasoning from the data already collected, Lara returned to the software and conducted an additional cycle with a cumulative Level 1 sample size. She used the relative frequency table for the outcomes of 36 rolls of the die as estimates shown on the right side of the poster (see Figure 3) for the “probability

after 36 trials.” This suggested that she had partitioned the problem and may have considered the task of estimating probabilities as separate from the task of conjecturing whether or not to refute the EPO hypothesis. Thus, their estimate of unequal theoretical probabilities did not correspond with a conclusion that all outcomes were equiprobable. Another interpretation was that they interpreted Question 3 (see Figure 2) as a request to provide relative frequencies of empirical data.

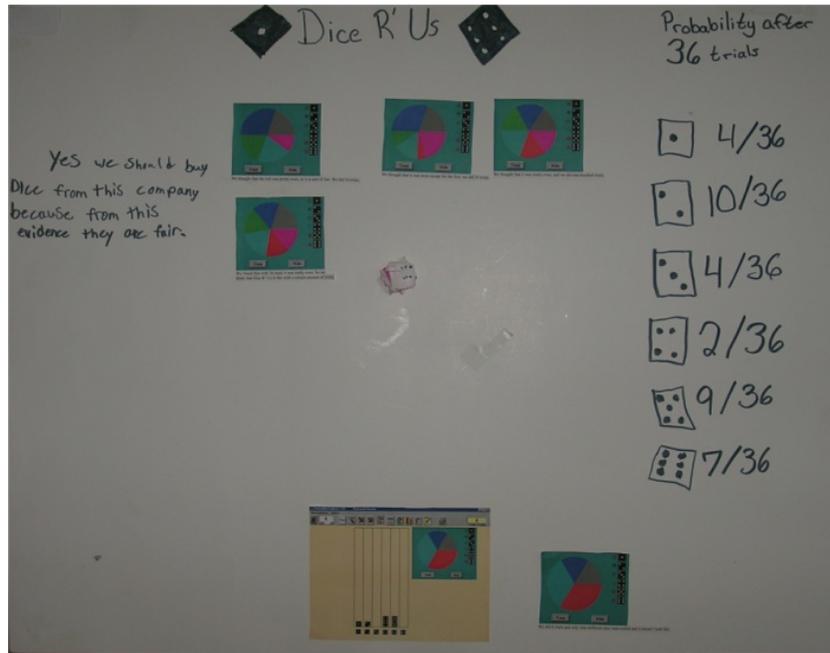


Figure 3. Dannie and Lara's poster

**Pair 2: Brandon and Manuel (High Rollers, Inc.)** Brandon and Manuel conducted eight cycles, 50% of which had sample sizes of 100 or below (see Table 2). For all but two cycles, they had at least three representations visible both during and after data collection: bar graph, pie chart, frequency table, or relative frequency table. Thus, they often had two representations of the same information available to them (e.g., the pie graph and relative frequency table) as well as representations showing different information (e.g., pie graph and bar graph). They predominantly reasoned about variability within a cycle, with two notable exceptions of reasoning across data sets.

Brandon and Manuel initially began by using the Run Until tool to test the number of trials before the first occurrence of a one. However they quickly abandoned this approach and conducted small trials which were added to the data previously collected. After 45 cumulative trials, they made an initial conjecture refuting the EPO hypothesis (“this is not fair”). To support this conjecture they both referred to the small number of fives that occurred. They explicitly reasoned about the variability within these data to support their development of a new hypothesis of the probabilities for the die outcomes, in particular that the five outcome seemed to have a smaller probability. However, they also were aware of the need to gather more data (“We need more evidence to conclude the company may be not fair.”).

In the next cycle, the students explicitly enacted their search for more evidence by conducting a Level 3 sample (500 trials). Although their first conjecture was refuting the EPO assumption, Brandon began comparing the new empirical data from the data collection process against his expectations. He appeared to use both an expectation formed in the last cycle (low probability of

five) and the initial expectation based on EPO to interpret the data in Cycle 2 (low number of fives and ones).

- Brandon: Interesting I have discovered that it is all fair but ... [after about 50 trials with all outcomes having frequencies of 6-11, and the “five” only occurring once]
- Manuel: Nuhuh! If you say that’s fair, you’re crazy. [points to the five column to support his argument]
- Brandon: I said it’s all fair except for five, that’s what I was going to say [now 110 trials and only 10 fives]... But anyway it’s doing pretty well. [Manuel scrolls to look at percents column with about 170 trials]
- Brandon: Dang 12 percent. [sings several times] [five is about 12%, but so is 1]

This coordination of two expectations illustrated his ability to attend to the dynamic simulation of trials and how that attention can begin to alter his hypothesis of the probabilities. After the trials were complete, he made a more explicit statement about how these data could be used to infer a new expectation and the implications of a non-uniform probability distribution (“So if you don’t want to get ones or fives, we are the company to be”). However, Brandon’s attention to variability within the data set did not seem to convince him of this new image (“Still, it’s pretty close to being fair, but it isn’t”).

Brandon’s tentativeness was apparent in the next cycle when, with a lower sample size (50), he returned to accepting that the data matched his initial hypothesis of EPO by comparing the variability between this set of data and their first set of 45 trials. In the current sample the frequency of fives was higher. To him, the differences in the variability across sets of data indicated there was no pattern and he made a conjecture of “fair,” and placed a “bet” against Manuel who was still unsure. However, in the next two cycles of Level 2 sample size, they appeared to be in agreement that the data represented equally likely outcomes and they made a written conjecture that “we have concluded from the info we have collected so far that the dice is fair.” Their statement indicated that up to this point in data collection, with a variety of small and large number of trials, they believed that the evidence was supporting a conjecture to confirm an EPO hypothesis.

While conducting 300 trials (Cycle 6), they continually tried to “cheer” the data into matching their mental image of EPO (e.g., “Come on! Get even, five!” and “Well, [five] just got off to a bad start”). Brandon noticed the variation within the data — ones, threes, fives, are a “bit behind” — but was still unwilling to reject the notion of EPO. When asked by the teacher/researcher to support their conjecture of “fair,” Manuel reasoned about the variability between this data set and the previous one (“The last time three and two were high.”). Manuel’s image of EPO was strengthened by his attention to random variation *across* data sets, while exhibiting a high tolerance for variability *within* the data (“they all [outcomes] didn’t have to be the same...it was just the luck”). Manuel’s high tolerance for variability was similar to that found in other studies (Lidster et al, 1995; Shaughnessy & Ciancetta, 2002) and his notion of “luck” was aligned with prior work that showed that students account for randomness of a die with references to luck (e.g., Lecoutre, 1992; Pratt 2000; Watson & Moritz, 2003). Brandon continued to question his image of EPO by his attention to variability *within* a data set and noticing a stabilization of certain outcomes occurring less often (Drier, 2000a; Parzysz, 2003). Interestingly, Brandon did not try to convince Manuel until the sample size reached above 1000.

- 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, we’re 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 frequencies in bar graph].
- Manuel: It's not that unfair. See, three and five and one are practically the same.
- Brandon: So they [three, five, and one] must have the same probability but that [points to relative frequencies for two and four] might have been more because look these [points to two, four, six] are always a bit higher than these [points to one, three, five].
- Manuel: [interrupting] Wait a second. Wait a second. We *are* unfair. These two, all these [points to one, three, five] are one and these [points to two, four, six] are two [referring to hypothesized weights in Weight Tool as he points to frequencies]. So, two, four, six [pause]... We're unfair.

Their social interactions helped facilitate both of them making connections between the empirical data and the theoretical probability as they coordinated the role of sample size, the distinct pattern in the variability within the data set, and the rearranging of their mental images to incorporate this new information. This was evidenced by Manuel's revelation that the die was "unfair" by Brandon's reference to the empirical distribution implying equal probabilities for outcomes one, three, and five and that these are less than the probabilities for two, four, and six.

During their last cycle with a Level 5 sample size, the teacher/researcher asked them to explain how the data could help them estimate the probabilities for each outcome (see Question 3 in Figure 2). Manuel responded "these percentages help us in, like, determining what the probability is moving to" and further explained how he was mapping each of the relative frequencies from the empirical data onto a model for the theoretical probability. His explanation was similar to the reasoning described by Prodromou (2007) of how students can use a data-centric distribution to inform a model-centric distribution. Manuel was having this conversation while data were continuing to be collected and the values in the table and graphical displays were being updated. In this regard, he seemed to appreciate the stabilization of the relative frequencies as an indicator of the theoretical probabilities. He further commented that he could not make a definitive conjecture about the probabilities as "we aren't sure which [probability to assign] depending on the outcome of this humunga [sic] big trial." His comment indicated an appreciation that, although there is a stabilization of the results, there will be slight changes in those results in future trials. Thus, he was demonstrating a strong conceptual understanding of the role of the law of large numbers in using relative frequencies in empirical data to make an estimate for a probability distribution.

Brandon and Manuel's poster (Figure 4) further highlighted the refinements they made to their initial mental image of EPO while reasoning from empirical data. On their poster (Figure 4), they documented how they once asserted that the die was fair with a small sample and then how larger samples gave them more confidence to make a claim of unfair and to estimate the probabilities ("We have changed our minds, the dice are *so* unfair"). This "mind changing" is aligned with findings from Pratt (2000) and Pratt and Noss (2002) where students developed an understanding that an increase in number of trials resulted in a more stable empirical distribution.

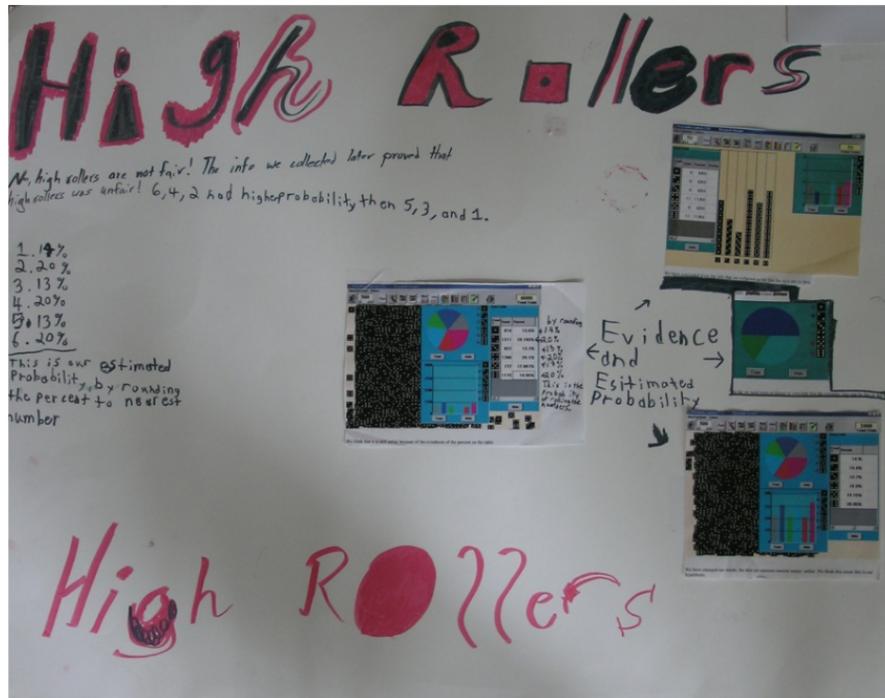


Figure 4. Brandon and Manuel's poster

**Pair 3: Greg and Jasyn (Slice n' Dice)** 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). For the majority of their cycles, they had three representations, either during or after data collection, most commonly the bar graph, pie graph, and frequency table. Thus, they simultaneously utilized representations with complementary functions (Ainsworth, 1999). Greg and Jasyn never utilized the relative frequency table. In approximately 45% of their cycles, they explicitly reasoned about variation among outcomes within a distribution.

Based on their work in the first cycle, they used the Run Until tool to conduct trials until each outcome appeared at least once. They suggested this may indicate that the EPO hypothesis was valid (Jasyn: "We know there are all the colors [in the pie graph],"  
Greg: "What? It's even."). We inferred that their strategy and comments reflected their mental image of the theoretical distribution of EPO and the resulting implication for empirical data, that is, that each outcome should be able to occur.

Because of the large variation in the theoretical distribution (see Table 2), the Level 4 sample size ( $n = 561$ ) in Cycle 2 allowed Greg and Jasyn to notice the drastic variation *within* the empirical distribution (see Figure 5). As Greg noted, "It's five and it's only 20 [referring to the frequency of fives] the rest [other frequencies] are like way above." They in turn confidently refuted the EPO hypothesis ("Five is way too low to be a fair dice") and formed a new hypothesis of a probability distribution with a low number of fives. In essence, with this set of trials, they were able to observe a distinct pattern (stabilization) in their results (Drier, 2000b; Parzys, 2003). They did not, however, make an explicit inference about how the low number of fives was linked to the probability of five occurring, as opposed to Brandon and Manuel who hypothesized the weights (probability) needed in the Weight Tool.

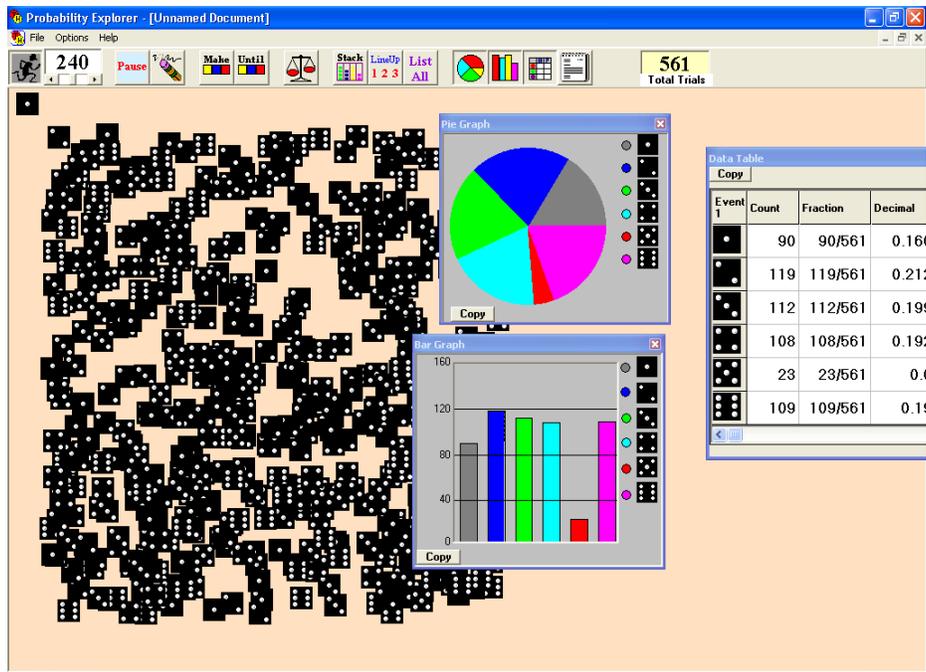


Figure 5. Greg and Jasy's screen display after 561 trials

Their new hypothesis that the die tended to produce few fives seemed to influence their subsequent strategy for collecting empirical data in Cycles 3-11. Thus, it appeared they were attempting to further substantiate their conjecture to refute the EPO assumption (“Okay this dice isn’t fair, we know that.”). One way they did this was their use of the Run Until tool to test whether it would take a substantial number of trials until the first occurrence of a five. With this tool, a sample size is not predetermined. They found that “it took 42 tries to get a five!” and conjectured that the die was “unfair,” supporting their previous conjecture. Their surprise with this result and the decision they made was consistent with the small probability (0.000567) that fives would not occur until at least the 42<sup>nd</sup> roll of an equiprobable die.

Following this cycle, the remaining cycles had relatively small sample sizes. Their results supported their previous finding of too few fives and they concluded that the die was biased. In essence, their one large sample seemed to help them form a new hypothesis for a probability distribution that was consistent with their remaining trials. Although they did not discuss their reasoning for using lower sample sizes in subsequent cycles, the supporting data for their new distribution may not have compelled them to use larger samples. In fact, they may have been engaging in a more playful exploration after they felt convinced of the claim about an unequal distribution. However, as can be seen in the Appendix, in most cycles, they continue to comment on their results and make a claim of “refute.” This provides evidence that they were staying engaged in the task.

The teacher/researcher directly asked them to use the evidence from the empirical data to conjecture probabilities for each outcome. Although they explicitly used the bar and pie graphs to estimate the relative frequency (“it’s sixes, well that’s about one-quarter right there” [Gesturing to the pie graph]), their lack of use of the relative frequency table seemed to hinder their ability to estimate a probability distribution and to instead provide a frequency distribution based on a set of 100 trials (Figure 6).

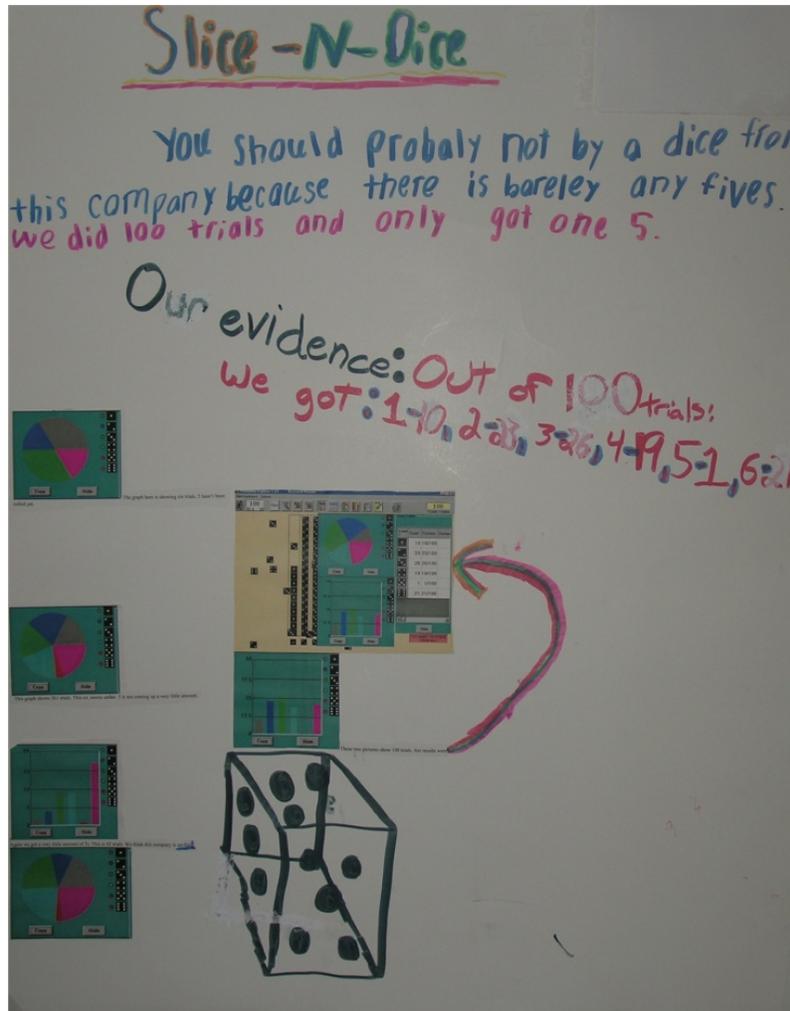


Figure 6. Greg and Jasyn's poster

The external resource of the task design for this pair (i.e., skewed bias in Slice n' Dice probabilities), combined with their use of a large sample size in Cycle 2 prompted a coordination of their initial (equally likely) and new (biased) images with the empirical data. This coordination was fostered by their attention to variability within data sets and invariance (low fives) across data distributions, and their explicit reference to various representations of data.

## 6. DISCUSSION

### 6.1 ENHANCING THE FRAMEWORK FOR INFORMAL HYPOTHESIS TESTING

Our main research question focused on how students were using the software to conduct their investigation, and specifically how they used the empirical distributions to confirm or refute a hypothesis. In each of the three pairs, students were coordinating their mental image of expected outcomes based on an imagined theoretical distribution (initially with the EPO hypothesis) with empirical data that either confirmed or conflicted with their image of the expected distribution. The initial categories that emerged from our data analysis included five types of decisions and actions the students engaged in during a cycle of investigation: 1) how to collect data (e.g., number of trials, adding data to previous trials), 2) whether to view representations (pie, bar

graph, data table, etc.) during or after the simulation, 3) whether to save the data (or not) as evidence for further analysis and justification, 4) whether to make a conjecture based on data to reject (or not) the initial hypothesis, and 5) whether more data were needed. The fourth and fifth decisions are already integrated in the framework at the point of an Interpretative Decision (see Figure 7). However, Figure 7 now includes the other decisions in the investigative cycle and a specific note of where students attend to variability. This enhanced framework may be useful for others in closely examining the tasks students engage in when using simulation tools in a statistical investigation, particularly with informal hypothesis testing.

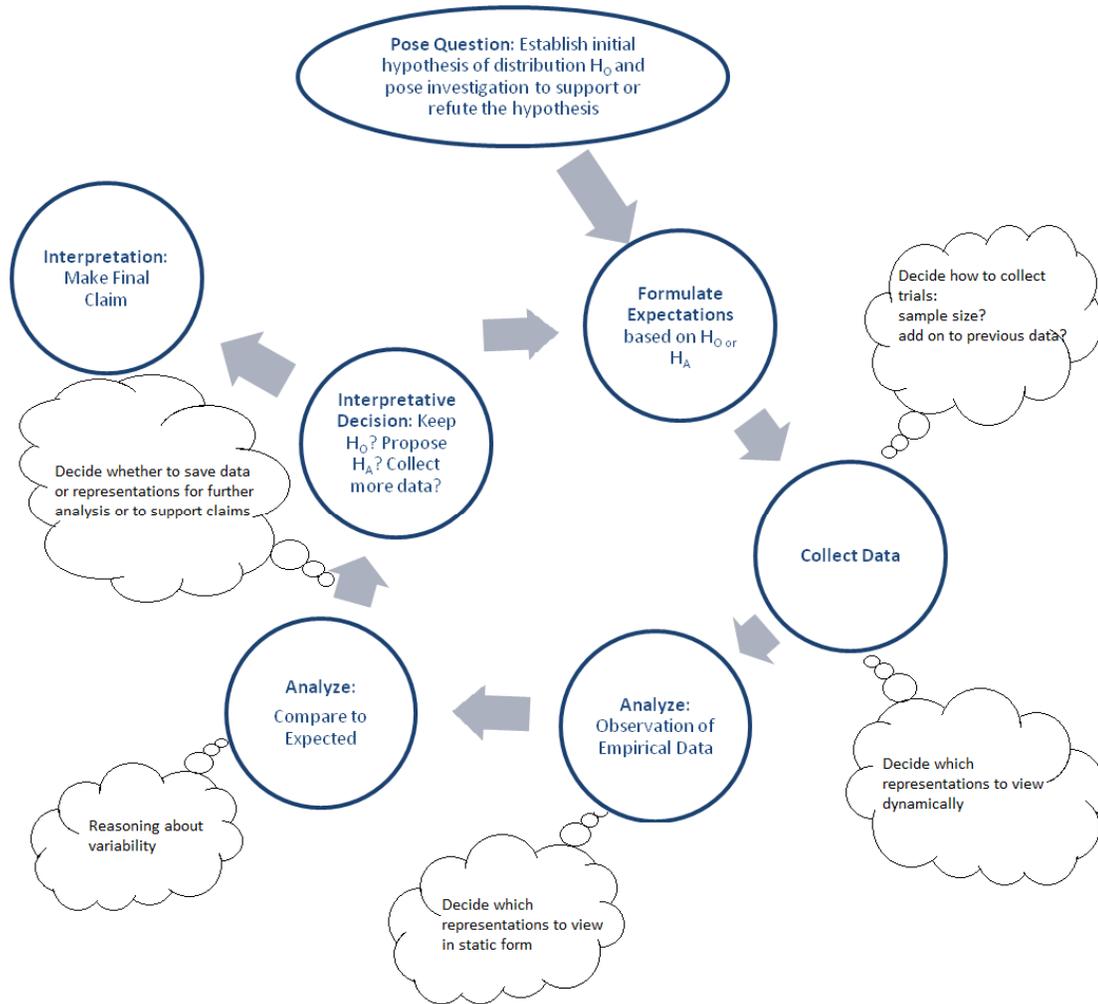


Figure 7: Investigative cycle for informal hypothesis testing with key decisions and reasoning

What also became apparent to us in the detailed analysis of the pairs' work in the *sequence of cycles*, was that students' images of a hypothesized probability distribution made in the *Formulating Expectations* phase (Figure 7) were consistently informing their:

- 1) decision of how to run trials (e.g., increase/decrease sample size, use of Run Until to test number of trials to get five, increase trials until data look even);
- 2) attention to results during data collection (e.g., cheering for five to catch up, noticing stabilization of relative frequencies with larger trials);

- 3) attention to results after data collection (e.g., one, three, and five are less than the others, comparing results to prior data sets); and
- 4) decision of whether the current data set should alter their hypothesis of the probability distribution (e.g., noting that they are unsure and need more data, accepting that data match their current hypothesis and expectations well enough, altering their current hypothesis of a distribution to account for new data).

Even though all pairs demonstrated this back and forth movement between their expectations and the process of collecting and analyzing empirical data to compare to this expectation, the robustness of the connections they were making was different. The way in which students decided to collect and display data (e.g., representations used) varied across the pairs, and thus the overall success of providing compelling evidence of whether or not to refute the EPO hypothesis and proposing a theoretical probability distribution was different.

Brandon and Manuel eventually made robust connections between empirical data and expected distribution both to refute the initial EPO hypothesis and to make an estimate for their new hypothesized distribution of probabilities. To accomplish this, they attended to several complementary representations of the data (Ainsworth, 1999), including the use of the frequency table after data collection much earlier than the other two pairs. They were also the only group to attend to the relative frequency table both during and after a simulation, with the exception of Lara who used it on the final run to estimate probabilities for inclusion on their poster. Brandon and Manuel referred to the relative frequencies often to justify their reasoning with each other and the teacher-researcher. The importance of the use of relative frequency displays of data was also noted by Ireland and Watson (2009) in helping students make a connection between empirical distributions and theoretical probabilities.

Thus, in considering the cycles portrayed in Figure 7 where one is continually having a mental “conversation” between observed data and expectations from an hypothesized distribution, the number of cycles is not what matters most (Brandon and Manuel utilized the least number of cycles), but well-connected and tight links that students develop between data and a model of a distribution. It is the careful analysis that students do that coordinates the two and includes attention to the role of sample size and variability.

## **6.2 REASONING ABOUT SAMPLE SIZE AND VARIABILITY**

Our second research question focused on how students attend to sample size and variability during their investigation. Because students had autonomy in selecting their sample size, it is interesting to note that all pairs chose varied sample sizes rather than systematically increasing the number of trials. Although the students had participated in several other probability experiments in the instructional unit and had engaged in class discussions where classmates shared their observations of an “evening out” (stabilization) effect with large numbers of trials, their inconsistent use of large samples indicates that they did not necessarily expect the “evening out” phenomenon to be immediately helpful in the task of testing whether or not the dice are fair. Pratt and Noss (2002) observed that newly developed internal resources (noticing “evening out”) often have lower cueing priority in contexts other than that in which the internal resource was developed. Our findings support this observation and also agree with the results from Pratt and colleagues’ (2008) recent study where students did not seem to attend to the effect of sample size in the variations observed in empirical distributions.

Greg and Jasyn quickly used a large number of trials when they suspected that one of their outcomes might have been more difficult to obtain. However, their reasoning about the large sample included more of a focus on how few fives occurred rather than on a stabilization of relative frequencies. In the language of Pratt and colleagues (2008), they were more focused on a local perspective (one outcome occurring significantly less often than the other five outcomes) rather than on a global perspective (stability of entire distribution after a large number of trials).

Dannie and Lara as well as Brandon and Manuel did not attend to the importance of a large sample size until near the end of their data collection. Dannie had a goal to see when the results would appear even that likely took away from her ability to notice the stabilization effect as Lara did. However, with substantial argumentation and interactions, Brandon and Manuel were able to come to recognize the stabilization after a large number of trials as a way to connect the empirical data to a theoretical distribution.

Regarding variability, students' reasoning within or across data sets yielded different inferences. It is interesting to note that all case study pairs generally reasoned about the variability *within* a data set (e.g., "Five is beating one") rather than consistently reasoning *across* samples (e.g., "We got a lot more fives this time"). The availability of the table and bar graph seemed to facilitate these within-sample comparisons. Brandon, Manuel, Greg, and Jasyn's reasoning about variability within a sample seemed to emerge as they viewed the simulation as a "race" between the outcomes of the die (e.g., "Come on five, catch up!", "Five is a bit behind"). The dynamic nature of the technology may have contributed to this phenomenon and suggests that when using dynamic technologies students' data collection, analysis, and interpretive decision making (see Figure 7) may occur in an integrated manner rather than in distinct phases. The "cheering" phenomenon can also encourage "students to notice and interpret representations that provide new insight about the phenomena" (Enyedy, 2003, p. 378). Cheering may also focus students' attention on variability by "observing the fluctuation of samples ... and observing the stabilization of the frequency distribution of the possible outcomes" (Parzysz, 2003, p. 1) within a run of trials.

We observed incidences when students reasoned across data sets to refute the EPO assumption and at other times, to support their image of EPO. Because these incidences occurred with widely varying sample sizes, it appeared students were not always coordinating their attention to variability with the notion of sample size. We agree with Pratt et al. (2008) that what seems to be important in fostering students' reasoning across data sets is a coordination of noticing invariance (similarities) and variance (differences) with the role of sample size. That is, students should come to expect different amounts of variability across data sets with small and large samples.

Dannie never explicitly reasoned about comparisons across samples. Lara did so once when she noted that fives were not low in a current data set of 100 trials as compared to a prior one of 40, and that helped support her agreement with Dannie during that cycle that the die may be fair. Brandon used comparison of two low-sample-size data sets to justify that he saw no consistent pattern indicating which outcome appeared the least. Likewise, Manuel also used this reasoning to notice no consistent patterns; however, he was comparing a sample size of 300 with a prior one of 50. By the time Brandon and Manuel agreed to refute the EPO hypothesis, they seemed to focus on the current very large set of trials and the stabilization effect. Thus, they likely had no need at this point to compare across samples. Greg and Jasyn were the only students to explicitly use reasoning across data sets about invariance between distributions (five is always low). Perhaps their obviously biased die made it much easier for them to notice these consistent patterns and may suggest that if we want to promote attention to invariance across empirical distributions, then perhaps it can be productive to use contexts in which theoretical distributions are significantly different from a uniform distribution. This is aligned with reasons given by Konold and Kazak (2008) for not using uniform distributions in the contexts of their tasks posed to students.

### 6.3 SOCIAL AND TASK CONTEXT

Consistent with the findings of Lidster et al. (1995), some students believed that dice could be fair for "some numbers" and for "some trials." Some students also demonstrated a strong desire for their die to be fair, possibly because of an underlying belief about all dice having equally

likely outcomes (Green, 1983; Watson & Moritz, 2003), 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., many students in this study tended to overcome these beliefs as they obtained larger sample sizes and engaged in discourse with their partners.

Discourse played a key role in our students’ reasoning. At times we found evidence that the individuals had different images of a probability distribution while at other times, their interactions provided no evidence that their images were significantly different. Differences were evident when substantial disagreements between students occurred, often within cycles of larger sample sizes. These conflicts generally arose when one student was willing to accept a large amount of variation between outcomes for a fair die 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, Brandon and Manuel were able to reach consensus to refute the EPO hypothesis after their sample size became large, by using the empirical data and multiple representations to make arguments and justify their claim of “unfair.”

By way of contrast, Dannie and Lara agreed upon language (e.g., “pretty fair”) that reflected the relative non-commitment by Lara to agree with Dannie’s support of the initial EPO hypothesis. However, neither student made strong data-based arguments to support their claim. Greg and Jasyn, analyzing the most blatantly-biased die, never conflicted on the issue of refuting the EPO hypothesis. Instead, their social interactions seemed to center more on control of the mouse and which representations to view and save. We suspect that this lack of negotiation was largely due to their early Level 4 sample that showed a bias in the die. Thus, we suggest not having such obvious bias in tasks in future teaching and research contexts.

## 7. CONCLUDING REMARKS

Our research has helped to answer the call (Jones, 2005; Jones et al., 2007) for more research on students’ understanding about the connection between relative frequency of empirical data and theoretically-derived probability, and the use of technology tools in learning probability. Given an engaging task and open-ended software tools, students in this study were able to reason from the empirical data and make meaningful connections back to a model of the probability distribution. 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 observed data and a probability distribution. Although we used social interactions as an external resource in our analysis, we did not focus on the intricacies of these interactions and the collaboration process. A follow-up analysis with a focus on argumentation and social interaction would further illuminate how students’ interactions affected their individual understandings. An example of such an analysis of a different group of students’ work on the Schoolopoly task can be found in Weber, Maher, Powell, and Lee (2008).

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 are not subjective opinions based on experiences or intuitions, but instead are calculated from large samples of available data. There is a need for students to experience how such probabilities are derived, and to understand that they represent an estimate of underlying probabilities and conditions that are unknown to us as external observers. In addition, we believe the teaching of probability and statistics should foster an understanding of how one can notice a phenomenon, make an initial conjecture or hypothesis, design an

experiment to collect sufficient and appropriate data, and use data-based reasoning to justify whether or not to refute the initial hypothesis. Such experiences can provide fundamental reasoning and skills that will be useful in students' data-rich lives.

Although the Schoolopoly task does not have the same real-world importance as medicine, we nevertheless found the problem context to be engaging to students and one that yielded rich classroom discourse. We ponder whether the context of dice as familiar objects in students' game-playing experiences added too much complexity in students' ability to reason about relative frequency and a theoretical probability distribution. It would be beneficial to study students' work on 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 and others also promoted by Zieffler et al. (2008) may offer students opportunities to grapple with numerous issues central to the study and understanding of informal inference. 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 tasks and simulation tools (technological and hands-on) in a context in which social interactions are fostered can help them make sense of the independent nature of random phenomenon, the importance of sample size, and the need to attend to variability. A robust understanding of these three concepts—randomness, sample size, and variation—seems to play a critical role in students' understanding of a bi-directional relationship between empirical data and expectations based on a theoretical probability distribution.

## REFERENCES

- Abrahamson, D., & Wilensky, U. (2007). Learning axes and bridging tools in a technology-based design for statistics. *International Journal of Computers for Mathematics Learning*, 12(1), 23-55.
- Ainsworth, S. E. (1999). A functional taxonomy of multiple representations. *Computers and Education*, 33(2/3), 131-152.
- Batanero, C., Henry, M., & Parzysz, B. (2005). The nature of chance and probability. In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 15-37). The Netherlands: Kluwer Academic Publishers.
- Canada, D. (2006). Elementary pre-service teachers' conceptions of variation in a probability context. *Statistics Education Research Journal*, 5(1), 36-63.  
[Online: [http://www.stat.auckland.ac.nz/~iase/serj/SERJ\\_5\(1\)\\_Canada.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ_5(1)_Canada.pdf)]
- 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* (Vol. 2, pp. 691-696). Tucson, AZ.
- 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.
- Friel, S. N., O'Connor, W., & Mamer, J. D. (2006). More than "meanmedianmode" and a bar graph: What's needed to have a statistical conversation? In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook* (pp. 117-137). Reston, VA: National Council of Teachers of Mathematics.
- 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.
- Green, D. R. (1983). Shaking a six. *Mathematics in School*, 12(5), 29-32.
- 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.  
[Online: [http://www.iejme.com/032009/IEJME\\_p09\\_Ireland\\_Watson\\_E.pdf](http://www.iejme.com/032009/IEJME_p09_Ireland_Watson_E.pdf)]
- Jones, G. A. (2005). Reflections. In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 375-379). The Netherlands: Kluwer Academic Publishers.
- Jones, G. A., Langrall, C. W., & Mooney, E. S. (2007). Research in probability. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 909-955). Charlotte, NC: Information Age Publishing.
- Kerslake, D. (1974). Some children's views on probability. *Mathematics in School*, 3(4), 22.
- Konold, C. (1987). *Informal concepts of probability*. Unpublished manuscript, University of Massachusetts.
- Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics. *Journal of Statistics Education*, 3(1).  
[Online: <http://www.amstats.org/publications/jse/v3n1/konold.html>]
- Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovation in Statistics Education*, 2(1).  
[Online: <http://repositories.cdlib.org/uclastat/cts/tise/vol2/iss1/art1/>]
- Lecoutre, M. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics*, 23(6), 557-568.
- Lee, H. S. (2005). Students' reasoning with small and large trials in probability simulations. In G. M. Lloyd, M. Wilson, J. L. M. Wilkins, & S. L. Behm (Eds.), *Proceedings of the Twenty-Seventh Annual Meeting of the Psychology of Mathematics Education-North American Chapter*. Blacksburg, VA.  
[Online: [http://www.allacademic.com/meta/p24814\\_index.html](http://www.allacademic.com/meta/p24814_index.html)]
- Lidster, S. T., Pereira-Mendoza, L., Watson, J. M., & Collis, K. F. (1995, November). *What's fair for grade 6?* Paper presented at the Annual Conference of the Australian Association for Research in Education, Hobart.
- Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, 8(1), 82-105.  
[Online: [http://www.stat.auckland.ac.nz/~iase/serj/SERJ8\(1\)\\_Makar\\_Rubin.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1)_Makar_Rubin.pdf)]
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage Publications.
- 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.
- Parzysz, B. (2003, August). From frequency to probability. Some questions posed by the new French senior high school curricula. Paper presented at the 54<sup>th</sup> International Statistical Institute, Berlin, Germany.  
[Online: <http://www.stat.auckland.ac.nz/~iase/publications/3/3193.pdf>]
- Pfannkuch, M., & Wild, C. J. (2004). Towards an understanding of statistical thinking. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic Publishers.

- Powell, A. B., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *The Journal of Mathematical Behavior*, 22(4), 405-435.
- Pratt, D. (2000). Making sense of the total of two dice. *Journal of Research in Mathematics Education*, 31(5), 602-625.
- Pratt, D. (2005). How do teachers foster students' understanding of probability? In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 171-190). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal*, 7(2), 107-129.  
[Online: [http://www.stat.auckland.ac.nz/~iase/serj/SERJ7\(2\)\\_Pratt.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Pratt.pdf)]
- Pratt, D., & Noss, R. (2002). The micro-evolution of mathematical knowledge: The case of randomness. *Journal of the Learning Sciences*, 11(4), 453-488.
- Prodromou, T. (2007). Making connections between the two perspectives on distribution. In D. Pitta-Pantazi & G. Philippou (Eds.), *Proceedings of the Fifth Conference of the European Society for Research in Mathematics Education* (pp. 801-810). Larnaca, Cyprus: University of Cyprus.
- Prodromou, T., & Pratt, D. (2006). The role of causality in the coordination of two perspectives on distribution within a virtual simulation. *Statistics Education Research Journal*, 5(2), 69-88.  
[Online: [http://www.stat.auckland.ac.nz/~iase/serj/SERJ5\(2\)\\_Prod\\_Pratt.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Prod_Pratt.pdf)]
- Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal*, 5(2), 46-68.  
[Online: [http://www.stat.auckland.ac.nz/~iase/serj/SERJ5\(2\)\\_Reading\\_Reid.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Reading_Reid.pdf)]
- Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph and K. Carr (Eds.), *People in mathematics education* (Vol. 1, pp. 6-22). Waikato, New Zealand: Mathematics Education Research Group of Australasia.
- Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 957-1009). Charlotte, NC: Information Age Publishing.
- Shaughnessy, J. M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society*, Cape Town, South Africa. [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.  
[Online: [http://www.stat.auckland.ac.nz/~iase/publications/1/6a6\\_shau.pdf](http://www.stat.auckland.ac.nz/~iase/publications/1/6a6_shau.pdf)]
- Stohl, H. (2002). Probability Explorer (Version 2.01) [Computer software]. Durham, NC: Author.  
[Online: <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.
- Strauss, A., & Corbin, J. (1990). *Basic of qualitative research: Grounded theory procedures and techniques*. Thousand Oaks, CA: Sage Publications.
- Tarr, J. E., Lee, H. S., & Rider, R. (2006). When data and chance collide: Drawing inferences from empirical data. In G. Burrill (Ed.), *Thinking and reasoning with data and chance: 2006 yearbook of the NCTM* (pp. 139-149). Reston, VA: National Council of Teachers of Mathematics.
- Watson, J. M. (2009). The influence of variation and expectation on the developing awareness of distribution. *Statistics Education Research Journal*, 8(1), 32-61.  
[Online: [http://www.stat.auckland.ac.nz/~iase/serj/SERJ8\(1\)\\_Watson.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1)_Watson.pdf)]
- Watson, J. M., & Kelly, B. A. (2004). Statistical variation in a chance setting: A two year study. *Educational Studies of Mathematics*, 57(1), 121-144.

- 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(4), 270-304.
- Weber, K., Maher, C., Powell, A., & Lee, H. S. (2008). Learning opportunities from group discussions: Warrants become the objects of debate. *Educational Studies in Mathematics* 68(3), 247-261.
- Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, 7(2), 40-58.  
[Online: [http://www.stat.auckland.ac.nz/~iase/serj/SERJ7\(2\)\\_Zieffler.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Zieffler.pdf)]

HOLLYLYNNE STOHL LEE  
Department of Mathematics, Science, and Technology Education  
North Carolina State University  
Campus Box 7801  
502D Poe Hall  
Raleigh, NC 27695

**APPENDIX: CONDENSED CODES DESCRIBING CYCLES  
OF DATA COLLECTION FOR CASE STUDY PAIRS**

Dannie & Lara (Dice R' Us)					Brandon & Manuel (High Rollers)					Greg & Jasyn (Slice n' Dice)				
Cyc	S	RP	C	V	Cyc	S	RP	C	V	Cyc	S	RP	C	V
1	1	D: S, P A: S, P	U	WA2	1	2	D: S A: S	R	WU2	1	1	D: S, P A: S, P		
2	1	D: S, P A: S, P	U	WA2	2	3	D: S, FT, RFT, P A: P, FT	U	WU1 WA1	2	4	D: B, P, FT A: NONE	R	WU2
3	1	D: S, P A: S, P	U	WU2	3	2	D: S, FT A: S, FT	U	AAD C WU1	3	1	D: B, P, FT A: B, P	R	AAS
4	1	D: S, P A: S, P	U	WU2	4	2	D: S, B, RFT A: S, B, RFT	C		4	2	D: B, P, FT A: B	R	AAS WU2
5	2	D: S, P A: S, P			5	2	D: P, FT, B A: P, FT, B	C	WA2	5	2	D: B, P, FT A: P	R	
6	2	D: S, P A: S, P			6	3	D: P, FT, B, RFT A: P, FT, B, RFT	C	WA2 AAD	6	1	D: B, P, FT A: NONE		
7		D: NONE A: P	C	AAD	7	5	D: P, FT, B, RFT A: P, FT, B, RFT	R	WU1 WA1 WU2	7	1	D: B, P, FT A: FT		
8	1	D: S A: S, P	U	WA2	8	5	D: P, FT, B, RFT A: P, FT, B, RFT	R	WU2	8	1	D: B, P, FT A: FT	R	WU2
9	2	D: P A: P	U	WA2						9	2	D: B, P, FT A: FT	R	
10	3	D: P A: P	C	WA1						10	1	D: S A: FT, P, S	R	WU2
11	1	D: P A: P, FT	C	WA2						11	2	D: FT A: P, B, FT	R	WU2
12	1	D: P, FT A: P, FT												
13	1	D: P, FT A: P, FT	U											
14	1	D: P A: P												
15	3	D: P A: P		R										
16	3	D: P A: P		R										
17	4	D: P A: P, FT	UR	WU1					*					
18	4	D: P A: P, FT	CU	WU1					*					WA1
19	1	D: NONE A: P												
20	1	D: NONE A: RFT												

Cyc = cycle

S = Sample Size: Level 1 (1-40), Level 2 (41-100), Level 3 (101-500),  
Level 4 (501-1000), Level 5 (1001+)

RP = Representation Used

D= During data collection A=After data collection

S=Stacked Icons, B=Bar Graph, P=Pie Graph, FT=Frequency  
Table, RFT=Relative Frequency Table

C = Conjecture

C= Confirm assumption of Equally likely outcomes (e.g.  
“that’s fair, etc)

U = Unsure (e.g. “sort of fair”, “not sure yet”)

R = Refute assumption of Equally likely outcomes (e.g. “no  
way”, “we are so unfair”

V= Variability

WA1, WA2 = Reasoning within data set, accepting of  
variability, # of students agreeing

WU1, WU2 = Reasoning within data set, unaccepting of  
variability, # of students agreeing

AAD= Reasoning across data sets, attending to deviations

AAS= Reasoning across data sets, attending to similarities

\*Disagreement between students that never reached consensus