ARGUMENT STRUCTURES OF RESPONSES TO A CONTEXTUALLY PROVOCATIVE HOSPITAL PROBLEM VARIANT

RANDALL E. GROTH
Salisbury University
regroth@salisbury.edu

JAMES P. BARRY
Salisbury University
jpbarry@salisbury.edu

ABSTRACT

Numerous variants of Kahneman and Tversky's (1972) hospital problem have been used to investigate intuitions about the Empirical Law of Large Numbers (eLLN). A largely separate line of research has focused on interactions between context knowledge and statistical reasoning. The present study merges these two lines of research by analyzing tertiary students' reasoning about a hospital problem variant set in a provocative context from their academic major. Participants' reasoning structures were diagrammed and compared against one another. Some responses closely matched an anticipated argument structure, and others differed along dimensions such as syllogistic structure, types of justifications offered, and task interpretation. Results of the study illustrate the importance of going beyond the metric of participants' success rate choosing the intended sample when doing research with contextually provocative hospital problem variants. Analyses of responses to such variants can be enhanced by examining the depth of eLLN intuition they reflect, their underlying syllogistic structures, and the extent to which application of the eLLN is qualified as needed in a given context.

Keywords: *Statistics education research; Argumentation; Context; Empirical Law of Large Numbers; Probability; Statistics*

1. INTRODUCTION

A voluminous body of research stems from Kahneman and Tversky's (1972) "hospital problem" (Figure 1). The problem can be solved using the Empirical Law of Large Numbers (eLLN), which is the idea that "a large sample is better than a small sample for estimating a population parameter" (Sedlmeier & Gigerenzer, 1997, p. 35). The eLLN provides a basis for recognizing that a small random sample of a population is more likely to yield extreme results than a large one. Few participants in Kahneman and Tversky's original study appeared to recognize this sample–population relationship, as only 20% chose the smaller hospital for the item shown in Figure 1. These findings inspired researchers to further investigate the prevalence of eLLN reasoning using myriad variants of the hospital problem that featured different contexts, wording, numerical values, and answer choices (Lem et al., 2011; Weixler et al., 2019). An example of a hospital problem variant that differs from the original along all these dimensions (Tabor & Franklin, 2019) is shown in Figure 2.

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50% and sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?
-The larger hospital
-The smaller hospital
-About the same (i.e., within 5% of each other)

Figure 1. The original hospital problem (Kahneman & Tversky, 1972, p. 443)

In a recent NBA season, Klay Thompson made 47% of his shot attempts. In his first game of the following season, he made only 5 of his 13 shot attempts (38.5%). Assuming Thompson's ability to make a shot is 47%, which of the following performances is more likely? Explain your reasoning.

- Completing at most 38.5% of shots in 13 attempts.
- Completing at most 38.5% of shots in 130 attempts.

Figure 2. An example of a hospital problem variant (Tabor & Franklin, 2019, p. 51)

Research using hospital problem variants has produced a wide range of success rates for participants in choosing the intended sample, from as low as 7% to as high as 87% (Sedlmeier & Gigerenzer, 1997). In a review of hospital problem variant research, Lem (2015) observed that "reasoning processes are easily influenced, making it difficult to study which characteristics of tasks and participants make a difference" (p. 791). The specific ways in which reasoning processes are influenced by different variants can be difficult to discern in studies that focus mainly on the metric of participants' success rate in choosing the intended sample. In response to this problem, Weixler et al. (2019) called for more in-depth qualitative examination of participants' responses to supplement the extensive quantitative research that has already been done on success rates with hospital problem variants. The goal of the present study was to contribute such qualitative information to the literature by analyzing arguments given in response to a hospital problem variant set in a context of interest to participants. The specific research questions were: (i) What reasoning patterns are associated with participants' sample choices for the hospital problem variant?; (ii) How do participants use knowledge of problem context and the eLLN to reason about the variant?

2. TOULMIN'S MODEL AND HOSPITAL PROBLEM VARIANTS

Responses to hospital problem variants can be construed as arguments in favor of one of the answer choices (e.g., larger sample, smaller sample, or neither). Toulmin's (1958, 2003) argumentation model provides a means to characterize the components and structure of such arguments, and it has been profitably used in past statistics and mathematics education research (e.g., Chazan et al., 2012; Gil & Ben-Zvi, 2011; Groth & Choi, 2023; Knipping & Reid, 2015). Toulmin argument components include data, claims, warrants, backing, qualifiers, and rebuttals. The ways these components may appear in a successful response to the original hospital problem (Figure 1) are shown in Figure 3. Next, we use Figure 3 as an outline to explain Toulmin model components and structure and how the model applies to hospital problem variant responses.

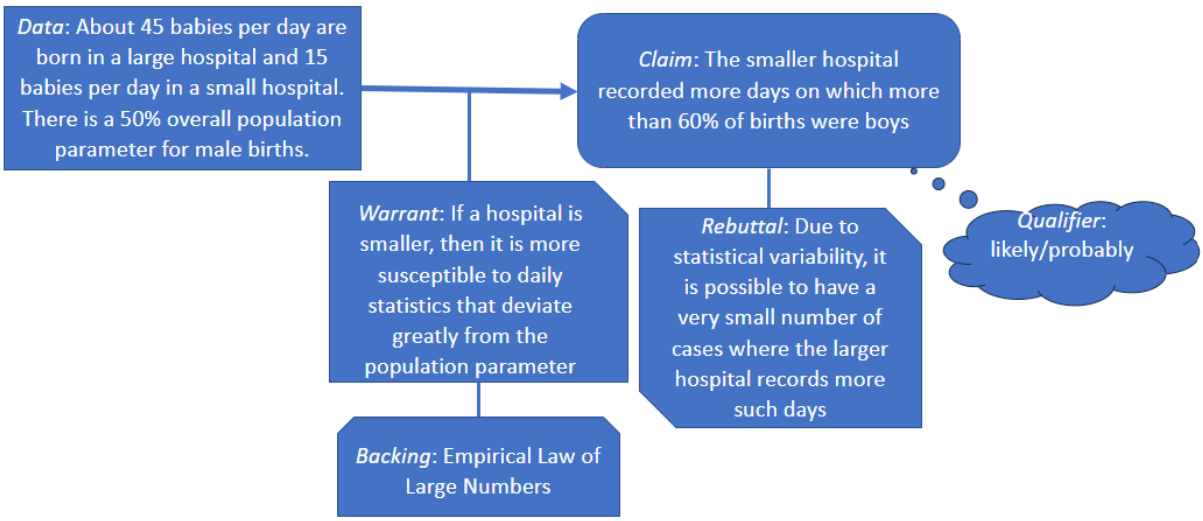

Figure 3. Toulmin diagram depicting the components and structure of a successful argument offered in response to the original hospital problem (Figure 1)

2.1. JUSTIFYING STATISTICAL ARGUMENTS: DATA, WARRANT, BACKING, AND CLAIM

Data, warrant, and claim are core components of an argument. Data refer to the information on which a claim is based. The warrant permits movement from data to claim (Toulmin, 1958, 2003). In the case of the original hospital problem (Figure 1), we are given the data that 45 babies per day are born in a large hospital and 15 per day in a small hospital. We are also given a population parameter of 50% boys. A successful response would contain the claim that the smaller hospital recorded more days on which more than 60% of births were boys. The intended warrant to allow for movement from data to claim is that if a hospital is smaller, it is more susceptible to days that deviate greatly from the population parameter of 50% males. The warrant might be explicitly stated, or it may remain unstated and implicit, as it is in many arguments (Warren, 2010).

In arguments, backing is sometimes provided to support the warrant (Toulmin, 1958, 2003). It is appropriate to back the warrant for a hospital problem response by using the eLLN. The eLLN helps justify the idea that statistics from large samples are more likely to be near their corresponding population parameters and that those from smaller samples are prone to being farther away. Expressions of the eLLN can take various forms. Brown (2019) described some possible intuitive expressions as swamping, size–confidence intuition, balancing, and sample–population ratio. Those who use swamping recognize that sample statistics from large samples are less influenced by extreme values than those from smaller samples (Well et al., 1990). Size–confidence intuition is the idea that larger samples are more likely to resemble the overall population and have sample statistics closer to population parameters (Sedlmeier, 1999). Individuals who use balancing reason that large samples provide more opportunities for extreme values, in either direction, to balance one another (Well et al., 1990). Sample–population ratio intuition is the idea that larger samples capture a larger portion of the population (Bar-Hillel, 1979). Any one or a combination of these intuitions can be used to provide eLLN backing for the warrant in a hospital problem response.

It is important to note that a degree of syllogistic reasoning is necessary to move from data to a warranted claim. Syllogisms consist of major and minor premises and a conclusion. In the original hospital problem (Kahneman & Tversky, 1972), the major premise in a successful response corresponds to the warrant that extreme variation between sample statistics and their corresponding parameters is most likely to occur in smaller samples. The minor premise corresponds to the data that the second hospital has smaller samples. From these two premises, the conclusion corresponds to the claim that the second hospital is more likely to have extreme variation between sample statistics and corresponding population parameters. Syllogistic reasoning of this nature is also needed for variants of the original problem. For example, variants that ask which sample would likely be closer to the population parameter (e.g., Well et al., 1990), rather than asking which sample deviates more, still require major and minor premises and a conclusion. In such cases, the major premise would be that larger samples tend to have sample statistics that match their corresponding population parameters more closely. The minor premise would be that the first hospital (or equivalent for the given variant) has larger sample sizes, leading to the conclusion that the first hospital will tend to have sample statistics that match the corresponding population parameter more closely. Although the Toulmin model cannot be reduced to classic syllogistic reasoning, syllogisms nonetheless are still integral to the model (Keith & Beard, 2008)

2.2. IDENTIFYING LIMITATIONS OF STATISTICAL ARGUMENTS: QUALIFIERS AND REBUTTALS

The Toulmin model also accounts for ways to identify limitations of arguments. Qualifiers express the degree of certainty with which a claim can be made. The need to account for sample–to–sample variability in hospital problem variants makes it necessary to qualify one's sample selection with appropriate degrees of uncertainty. Sample–to–sample variability and sample size can make a given sample more likely to occur, but the selected outcome cannot be guaranteed. Language that acknowledges variability (Makar & Confrey, 2004), rather than deterministic language, is thus in order when making arguments to justify one's solutions to hospital problem variants. For example, in Kahneman and Tversky's (1972) original hospital problem, the selection of the smaller hospital might

be qualified by saying that it is "more likely" for a sample of 60% boys to occur there (as opposed to simply stating that it will occur there). Such statistical claims can also be qualified with other words, such as "probably," "maybe," "tend to be," and "usually" (Ben-Zvi et al., 2012; Henriques & Oliviera, 2016). Answers to statistical questions, often reached inductively, must be qualified with appropriate degrees of uncertainty; this differs from mathematical proof, in which sound deductive reasoning leads to certainty when stating claims (Arnold & Franklin, 2021; delMas, 2004; Rossman et al., 2006). Qualifying a claim may naturally lead to identifying rebuttals (statements of cases under which the claim does not hold), which could be atypical samples that run counter to the claim. Including qualifiers and rebuttals in a response demonstrates awareness of the stochastic, rather than deterministic, nature of solutions to hospital problem variants.

2.3. USING CONTEXT KNOWLEDGE IN STATISTICAL ARGUMENTS

Minimal context knowledge is needed to construct the successful hospital problem response depicted in Figure 3. In the original hospital problem and many of its variants, context serves only as the cover story and little contextual knowledge related to the cover story is required to justify or qualify one's sample size selection. Consequently, the cover story was not among the task characteristics Lem (2015) identified as having a strong impact on participants' success rates in previous studies. Researchers who seek to assess participants' eLLN intuitions may consider variants that require minimal context knowledge to be optimal for their studies because such items are more broadly accessible and, hence, yield more statistically generalizable findings. Items with sterile contexts are also more likely to require strict reliance on the eLLN as backing rather than contextual considerations.

Although items that require minimal use of context knowledge may be desirable for some research purposes, such items are also limited in the extent to which they can be considered authentic assessments of statistical reasoning. As Cobb and Moore (1997) observed, "data are not just numbers, they are numbers with a context" (p. 801). One of the hallmarks of statistical thinking is to engage in "continual shuttling backwards and forwards between thinking in the context sphere and the statistical sphere" (Wild & Pfannkuch, 1999, p. 228). Statistical questions about a given context can motivate gathering and analyzing relevant data, and creating plausible interpretations of data requires revisiting the context to make sense of why a given result was obtained and what it may mean. Traversing between data and context can ultimately generate new knowledge of the context (Bargagliotti et al., 2020; Ben-Zvi & Aridor-Berger, 2016). Reconciling data and context is a core component of empirical inquiry.

In some studies, students have been prompted to traverse between data and context during extended statistical investigations. For example, Langrall et al. (2011) had students analyze authentic data from engaging contexts and found that some students used context knowledge to justify and qualify claims about the data. Context knowledge allowed these students to offer personal opinions and logical arguments as justifications, and it also helped them identify limitations of their own arguments. Similarly, Shaughnessy and Pfannkuch (2002) described how students' context knowledge of the Old Faithful geyser helped them justify data-based predictions about the timing of future eruptions. Attaining success coordinating data and context in such a manner can, however, be challenging. Langrall et al. (2011) found that context knowledge at times led to extended discussions that did not help students move forward on data analysis tasks. Similarly, Pfannkuch (2011) found that students' inventive stories about unusual heights in another data set made it difficult for their teacher to focus attention on statistical ideas that were relevant to analyzing the data at hand. Moreover, students' contextual beliefs can be resistant to change. For example, Masnick et al. (2007) found that inaccurate beliefs about pendulum motion did not change even after students analyzed data that contradicted their beliefs. Such studies indicate that traversing between data and context requires managing one's intuitions about a situation, leveraging those that are helpful and setting aside those that are not.

The intentional management of one's intuitions about a situation requires suspending the natural proclivity to act immediately on them. In psychological terms, it requires System 2 thinking rather than System 1 thinking (Kahneman & Klein, 2009). Kahneman (2011) characterized System 1 thinking as "thinking fast", and System 2 thinking as "thinking slow," stating, "System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control" (p. 20), and "System 2 allocates attention to the effortful mental activities that demand it" (p. 21). Although System 1 thinking is vital for daily tasks such as recognizing faces and objects, detecting signs of danger, and reading common

words at first sight, it should give way to System 2 thinking in situations that require careful, systematic analysis. Reconciling statistical and contextual knowledge requires System 2 thinking if the task at hand merits more than automatic application of an intuition or mathematical procedure to produce a solution.

At times, System 1 thinking may lead students astray on hospital problem variants. Kahneman (2011) attributed high failure rates on the original hospital problem to participants' use of System 1 thinking that included an incorrect "law of small numbers" intuition, which led some to put too much faith in small sample sizes. Some schemas developed in the context of school mathematics that become part of students' System 1 thinking may also lead them to conclude that the two choices in hospital problem variants are equally likely. Fischbein and Schnarch (1997), for example, found that older students were more likely than younger ones to believe the two outcomes in hospital problem variants have the same probability of occurring. Older students tend to have more developed proportional reasoning schemas, which may be automatically activated when they see equal percentages or fractions in a hospital problem variant (Tirosh & Stavy, 2000). Although proportional reasoning schemas are helpful for a variety of other tasks, their automatic, incorrect application to hospital problem variants may occur after students have encountered many equal ratio situations in school mathematics exercises and consequently perceive the variant to be the same type of exercise (Sommerhoff et al., 2023).

It should be noted, however, that System 1 thinking is not always disadvantageous when responding to hospital problem variants. Stanovich (2018) noted that helpful normative intuitions can enter System 1 thinking and be deployed when warranted. The eLLN is one such normative intuition (Sommerhoff et al., 2023). Take, for example, a mathematics education researcher who has become an expert in posing hospital problem variants and analyzing students' responses to them. This researcher will often automatically recognize problems to which the eLLN applies and attain a solution without much effort. Similarly, students who have succeeded in responding to many such variants can quickly recognize many situations in which they should deploy the eLLN rather than an equal ratios schema. For such individuals, System 1 thinking suffices for most hospital problem variants. They would need to activate System 2 thinking, however, if there is any question about whether the eLLN is applicable in a variant's context.

3. METHOD

Because traditional hospital problem variants tend to be solved (either correctly or incorrectly) with System 1 thinking, to this point, hospital problem variant research has shed minimal light on how students coordinate statistical and contextual knowledge in settings in which the eLLN can be applied. However, there is evidence to indicate that some hospital problem variants do activate participants' context knowledge in different ways. Maxara and Biehler (2010), for example, found substantial differences in how participants responded to one hospital problem variant set in a casino context and another in a political survey context. Weixler et al. (2019) found quantitative differences in performance by gender on a hospital problem variant set in the context of tossing coins. To gain a deeper understanding of the possible use of context knowledge in such situations, we qualitatively investigated participants' responses to a contextually provocative (Madden, 2011) hospital problem variant; that is, an item designed to prompt multiple respondents to actively use context knowledge to construct arguments about their sample size choices. Focusing on argument structures elicited by a contextually provocative item positioned us to contribute to both the growing literature on the role of context in statistical reasoning (Nilsson et al., 2018) and the existing body of research on hospital problem variants (Lem et al., 2011). To increase the probability that participants' context knowledge would be activated when solving a hospital problem variant, we designed the study around a variant with a context (sports) from participants' academic major (physical education).

3.1. HOSPITAL PROBLEM VARIANT

The hospital problem variant for the present study is shown in Figure 2. It was from a textbook on statistical reasoning in sports (Tabor & Franklin, 2019). The variant asks respondents to assume that a basketball player has the ability to convert 47% of the shots he takes. Participants then had to decide if an unusually low conversion rate of at most 38.5% would be more likely in a small sample of 13 shots or a large sample of 130 shots and explain the reasoning underlying their sample size choice.

The basketball variant used for the study (Figure 2) merits thinking beyond that which is typical of System 1 thinking, even among those who are experts in solving hospital problem variants. The statistical analysis of basketball shots has provoked much past debate among statisticians. Statistical arguments have been provided both for and against the assumption that basketball shots are independent trials (Bar-Eli et al., 2006). Additionally, basketball shots potentially involve elements of motor skill, consistency, and affect that is normatively irrelevant to traditional hospital problem variants based on benign random processes such as flipping coins, rolling dice, and spinning spinners. Accordingly, we conjectured that the basketball variant would prompt participants to engage in more than just superficial use of their context knowledge or eLLN intuitions.

We also aimed to prompt participants beyond superficial System 1 use of mathematical knowledge to justify their arguments. As noted earlier, older students may opt for the "about the same" option (see Figure 1) because of the hasty application of an equal ratio schema related to similar school mathematics problems (Fischbein & Schnarch, 1997; Tirosh & Stavy, 2000). Rubel (2009) found evidence that some students who chose the "about the same" option for different variants actually leaned toward the larger or smaller sample when their thinking was probed further in interviews. One such student said he chose the "about the same" option because it was the "math test" answer. But, when asked to explain his answer choice, the student gave context-based reasons for favoring the smaller sample in "real life." In the task we used for the present study (Figure 2), the "about the same" option was left out to reduce the possibility of cueing automatic, superficial System 1 application of equal ratios. Because the variant required an explanation and not just a forced answer choice, participants could still express reasoning aligned with the "about the same" option, and one did. However, as reported later, many of our participants opted to bring context knowledge to bear in arguing for either the larger or smaller sample of shots.

3.2. PARTICIPANTS

The study's participants were 58 undergraduate physical education majors. According to university records, 43 were male, and 15 were female. All had taken or were currently enrolled in a course on the foundations of physical education. The course introduced the historical and philosophical foundations of physical education, fitness, and sport. It included the study of the cognitive, affective, and psychomotor domains of physical education (Society of Health and Physical Educators, 2013). The course prompted participants to draw on their knowledge of these domains to design inclusive and educative youth sports environments with Universal Design for Learning (Sherlock-Shangraw, 2013). Hence, the participants had academic backgrounds that could lead them to make conjectures about possible influences on an individual's athletic performance beyond just attributing observed differences to statistical variation.

Participants also had statistics experiences relevant to the variant used for the study. Their pre-college curricula were based on the Common Core State Standards in the United States (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), which includes ideas of sampling, probability, statistical variability, and informal inference. The Grade 7 Common Core includes "informal work with random sampling to generate data sets and learn about the importance of representative samples for drawing inferences" (p. 46). In high school, the Common Core states that students should "understand statistics as a process for making inferences about population parameters based on a random sample from that population" (p. 81). Additionally, 28 of the 58 participants had taken a college-level statistics course. The course was among the options for completing the university's general education mathematics requirement but was not required for the physical education major. The statistics course syllabus included random sampling, sampling distributions, confidence intervals, and hypothesis tests.

3.3. PROCEDURE

The hospital problem variant set in a sports context (Figure 2) was distributed to three classes taught by the second author of this article. Of the 78 undergraduates enrolled in the courses where the item was administered, 58 provided their consent to have their data used for the study. Approximately ten minutes were used during each class session to have participants write individual responses to the

problem and their explanations. We purposefully posed the problem as a physical education class activity rather than a mathematics or statistics class activity, conjecturing that this choice would further increase the likelihood of activating participants' context knowledge. We used a writing prompt format to encourage the deliberative System 2 thinking (Kahneman, 2011) needed to reconcile data and context because writing provides a space to examine and restructure one's thinking (Menary, 2007) rather than encouraging immediate verbal responses reflective of System 1 intuitions. The first author gathered students' written responses and, per the human subjects protocol for the study, removed participants' names and identifying information from them in preparation for collaborative analysis with the second author.

3.4. DATA ANALYSIS

A perspectivist approach to analysis (Cornish et al., 2013) was taken because the goal was to combine observations about the data that were made through two different disciplinary lenses. The first author was a researcher in statistics education, and the second author in physical education. Each analyst brought their own disciplinary lens to bear, in contrast to approaches in which experts from the same discipline calibrate their analyses to produce a homogeneous interpretation of the data. Having analysts with different disciplinary perspectives helped keep both the statistical and contextual aspects of participants' arguments visible in our characterizations of participants' arguments.

During the initial stage of data analysis, each author read participants' responses independently. The first author primarily attended to participants' use of statistics and mathematics, and the second author to participants' use of physical education context knowledge. During the initial independent readings, the first author wrote memos (Miles et al., 2019) on participants' use of statistical and mathematical ideas such as the eLLN, proportional reasoning, statistical variability, mathematical procedures, and calculations to explain their reasoning. The second author did the same for contextual elements of responses related to the domains of physical education the participants had studied (Society of Health and Physical Educators, 2013). This contextual lens made considerations from the psychomotor and affective domains visible, such as participant-conjectured effects of muscle memory, nervousness, confidence, fatigue, and practice.

To provide an initial analytic framework (Gerbic & Stacey, 2005) to organize the memos from the two initial independent readings of the data, the first author constructed the Toulmin diagram shown in Figure 4. It is a slightly modified version of the original hospital problem response structure shown in Figure 3 and approximates the reasoning that the textbook authors (Tabor & Franklin, 2019) intended to elicit with the basketball variant (Figure 2). It also includes a sample qualifier and rebuttal that would be appropriate for the argument. Although some participants' responses closely matched the intended structure, the majority did not. So, in the next phase of analysis, we collaboratively edited Figure 4 to represent the structures of participants' responses. This required adding, removing, and editing the node text in Figure 4 and changing connectors between nodes to depict relationships among elements within their arguments. Constructing Toulmin diagrams in this manner provided a means for us to keep the statistical, mathematical, and contextual elements of each response visible and to represent relationships among them. The Toulmin diagrams in the results section of this article represent our agreed-upon characterizations of the data after reviewing and editing multiple written drafts. As an additional step toward building the trustworthiness of our qualitative data analysis (Lincoln & Guba, 1985) in the results section of this article, several verbatim participant responses are provided (Eldh et al., 2020) alongside the finalized Toulmin diagrams to help readers further trace, assess, and replicate the reasoning processes that were used to produce qualitative characterizations of the data.

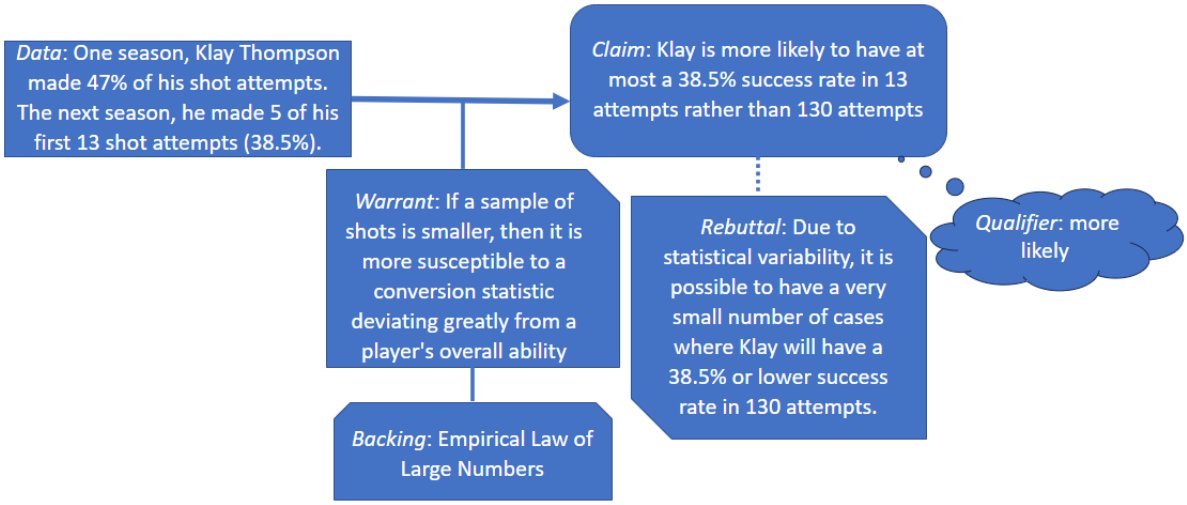

Figure 4. Toulmin diagram depicting the intended argument structure for a response to the hospital problem variant used in this study (Figure 2)

In the next stage of analysis, the Toulmin diagrams depicting participants' arguments were compared against one another and grouped according to similarities and differences in components and structure. First, we sorted the responses according to the type of claim they contained: smaller sample, larger sample, or neither sample. The first author further sorted arguments according to how closely they matched the syllogism, backing, and problem interpretation represented by the intended argument structure shown in Figure 4 and then grouped them into categories. The second author of the article audited the first author's qualitative categorization scheme (Lincoln & Guba, 1985) by comparing the category descriptions against the original data and the earlier collaborative analyses. The agreed-upon characterizations of argument categories were summarized using a matrix display (Miles et al., 2019). We revisited and refined the categories in the display after receiving feedback from the anonymous peer reviewers of a previous draft of this article, who offered alternative possible categorizations of sample data excerpts included in the draft. The final categorization of participants' arguments is shown in Table 1.

Table 1. Number of participants exhibiting each type of argument structure and their sample size choices

Response category	Sample size choice		
	Small sample	Large sample	Neither sample
Closely matches intended argument structure with eLLN backing	13	0	0
Uses eLLN backing but contains an invalid syllogism	0	1	0
Relies primarily upon mathematical backing	1	1	1
Relies primarily upon contextual backing	9	7	0
No explicit backing provided	6	6	0
Response based on an unintended interpretation of the problem	10	3	0

4. RESULTS

In responding to the hospital problem variant for the study (Figure 2), 39 participants claimed that it would be more likely for Klay to make at most 38.5% of his shots in 13 attempts rather than in 130

attempts; 18 participants selected the 130 attempts option; and one participant claimed there should be no difference between 13 and 130 shots. Hence, 39 of the 58 participants selected the intended sample of shots, which was a 67.2% success rate. Participants who had taken an introductory college statistics course chose the intended sample at a slightly higher rate than those who had not. Those who had completed the course had a 71.4% rate of doing so (20 out of 28), whereas 63.3% (19 out of 30) of those who had not completed the course chose the intended sample.

The number of participant arguments that closely matched the intended response structure was substantially lower than the number that chose the intended sample. Overall, 13 of the 58 participants offered arguments with data, claim, warrant, and backing like those in Figure 4. (Hereafter, these responses are referred to as "closely matching" the intended argument structure.) Participants who had taken an introductory college statistics course closely matched the intended argument structure at a slightly higher rate than those who had not. Among those who had taken such a course, 25% (7 out of 28) offered a response structure closely matching Figure 4. Among those who had not taken the course, 20% (6 out of 30) had a response structure closely matching Figure 4. The qualitative categories of arguments we observed are summarized in Table 1. One such argument type contained eLLN backing but included an invalid syllogism. Others relied primarily upon backing that was mathematical or contextual in nature. There were also cases where no explicit backing was provided, and some responses were based on unintended interpretations of the hospital problem variant. The numerical distribution of arguments into these different categories is shown in Table 1. Next, illustrative responses from each category are provided, along with descriptions of the reasoning we used to model their components and structures with Toulmin diagrams.

4.1. RESPONSES CLOSELY MATCHING THE INTENDED ARGUMENT STRUCTURE

In the set of participant responses, 13 (first row of Table 1) could be modeled with minimal changes to Figure 4 because they had the intended syllogistic structure and drew upon the eLLN as backing. In three of the 13, there was evidence that participants used multiple expressions of the eLLN to back their arguments for choosing the smaller sample of shots. S28, for example, wrote,

> In the case of Klay Thompson, if he has the ability to shoot 47%, then it seems very unlikely that after 130 shots he would still be at 38.5%. With this in mind when only shooting 13 shots and each miss could deduct 7% of your accuracy it doesn't seem too crazy that he could have missed an extra shot and it have dropped him so substantially seeing that he hasn't shot that much. Therefore with Klay Thompson's ability to shoot 47% it is a lot more likely that he shot 38.5% after only 13 shots not 130.

In this response, S28 noted that the success rate from a larger sample of shots is likely to be in closer proximity to a population parameter of 47%, which reflects a size–confidence intuition (Brown, 2019) expression of the eLLN. S28 also noted that smaller samples are more likely to stray from the parameter because each shot in a small sample of 13 would account for approximately 7% of the shot conversion statistic. This second type of eLLN backing suggests a balancing or swamping intuition (Brown, 2019) in that conversion statistics for larger samples would not be as heavily influenced by the outcome of a single shot. Backing of this nature can be modeled by replacing the Figure 4 node representing backing with two nodes, with each of the two nodes representing one of the two eLLN intuitions expressed in the argument.

The other 10 of the 13 participants who stated eLLN intuitions in arguments with the intended syllogistic structure used one, rather than multiple, eLLN intuitions as backing. Eight of the 10 used size–confidence intuition, stating that a larger sample of shots would more closely approximate a population parameter of 47%. For example, S34 wrote, "The first one (smaller sample of shots) because the more shots he takes, the more likely he can boost his shooting percentages so the second one should start to increase toward that 47% mark." Similarly, S16 wrote, "The more shots he takes will average out closer to his true ability to make shots which is 47%." Two of the 10 giving a single expression of the eLLN used a balancing or swamping intuition to back their arguments. S50, for example, claimed that completing at most 38.5% of shots in 13 attempts would be more likely than completing at most 38.5% in 130 attempts, writing, "The smaller the sample size the more likely it is for the data to be skewed since each individual shot means more."

As illustrated in the sample responses given above, some arguments contained qualifiers, and others did not. The three participants who offered multiple forms of eLLN backing also included appropriate qualifiers such as "very unlikely" and "a lot more likely" (e.g., S28). Four of the 10 who offered one form of eLLN backing included qualifiers, and the others did not. The use of language of uncertainty such as "more likely" and "should" (e.g., S34) to qualify an argument rather than deterministic language such as "will" (e.g., S16) is worth noting because it contributes to a more fully developed response to an inherently stochastic item. Rebuttals, as shown in Figure 4, can contribute further to a well-developed response, but they were not present among any of the responses in the participant group. Given these considerations, Figure 4 can be used as a tool to represent responses like those evident among the first 13 participants we have described by removing the rebuttal node and either removing or changing the text of the qualifier node to match the given response. Qualifier nodes can also be added as needed when there are multiple qualifiers in a response (e.g., S28).

4.2. RESPONSE WITH ELLN BACKING BUT AN INVALID SYLLOGISTIC STRUCTURE

The argument of one participant (second row of Table 1) who used a form of eLLN backing did not have a valid syllogistic structure, as its conclusion did not follow from its premises. S2 claimed that completing at most 38.5% in 130 attempts would be more likely than a 38.5% conversion rate for 13 attempts. To justify this claim, S2 wrote, "Because as shown in his first game, measuring only 13 attempts there is much more of a chance an unusual measure could happen, whereas if 47% is his true shooting percentage, it should easily average out to over 38.5% made." Although this response resonates with balancing and size–confidence intuitions of the eLLN (Brown, 2019), the backing contradicts the idea of moving to the claim that taking more shots will produce a conversion rate closer to 38.5%. Hence, although S2 appeared to use eLLN intuitions, the response contained a syllogistic difficulty in that the conclusion (selecting the larger sample) did not follow from the major premise (the eLLN). Given S2's demonstration of sound eLLN intuitions within an invalid argument structure, it is possible that the linguistic complexity of the problem (e.g., interpreting the meaning of "at most"), rather than flawed statistical intuition, caused the observed syllogistic difficulty. Because this type of syllogistic difficulty turned up in other categories of response as well, in Figure 5, we introduce the convention of placing an octagon on the path from data to claim to represent responses in which the warrant and/or backing run(s) counter to the claim, using the specific components of S2's response as an example.

4.3. RESPONSES WITH PRIMARILY MATHEMATICAL BACKING

In three cases, mathematical backing was offered, but the eLLN was not used (third row of Table 1). One respondent, S36, relied solely upon proportional reasoning as backing, explaining:

> It would be equally as likely for Thompson to complete 38.5% of shots in 13 attempts as it would be for 130 attempts because he would make 5 of 13 or 50 of 130. It's just 10 times as many shots.

S36 was the only participant to go outside the sample size choices provided to claim that the two sample size choices would be equally likely. This type of response is prevalent in research on hospital problem variants that include "equally likely" as a third option (Lem et al., 2011). Proportional reasoning, in isolation, supports such a response because it opens the possibility of viewing any given sample simply as a smaller-scale replica of the overall population. The argument structure itself is syllogistically valid, as the justification given supports movement from data to claim. Hence, the response can be diagrammed using conventions we have already introduced. However, applying principles from the field of probability and statistics (Kahneman & Tversky, 1972; Lem, 2015), the argument is not compatible with disciplinary norms because it contradicts the law of large numbers. So, it serves as an example of an argument that is valid but not true. Two of the three responses offered neutral mathematical backing. These included "added a zero to the second situation" (S44) and "He shot 13 attempts and made only 5 so 5 divided by 13 is 38.5%." (S13). The mathematical observations in these cases were correct but not enough to support a claim for the choice of a larger or smaller sample of shots.

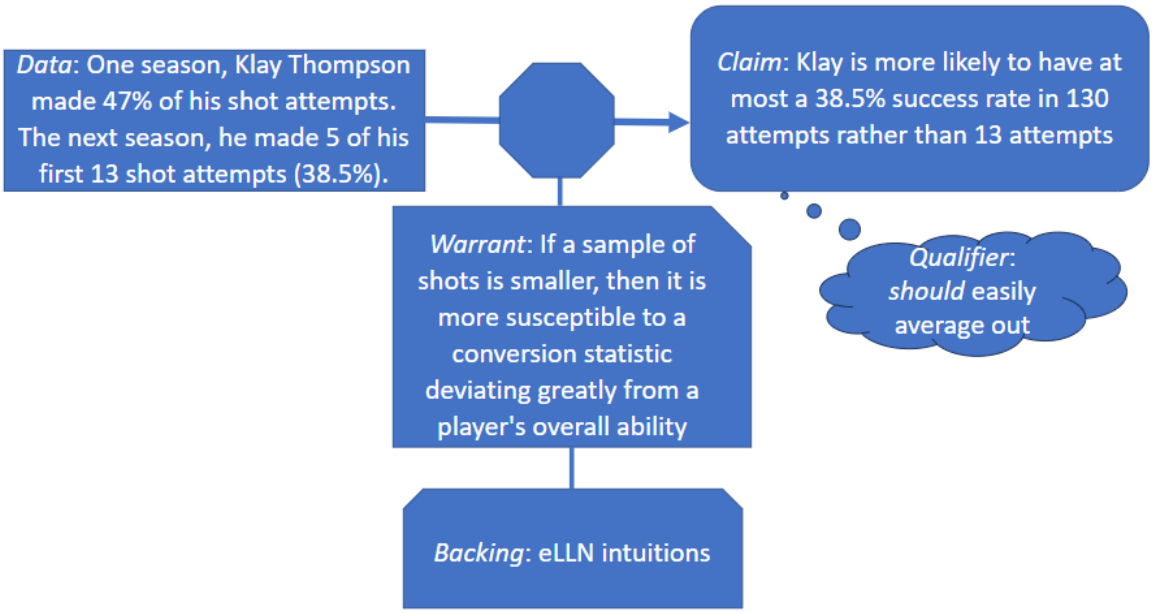

Figure 5. Representation of an argument backed by eLLN intuitions in which the justification provided runs counter to the movement from data to claim

4.4. RESPONSES WITH PRIMARILY CONTEXTUAL BACKING

Sixteen responses (fourth row of Table 1) relied primarily upon contextual backing without referencing eLLN intuitions. Four of these 16 participants suggested that a player's ability to make shots improves with the number of shots taken rather than being a fixed parameter. Five of the 16 offered observations about how shooting performance varies from one game to another. Four of the 16 focused primarily on observations of Klay's shooting skill to back their arguments. Three of the 16 backed their arguments using knowledge of how basketball conversion percentages are ordinarily determined. Some participants who relied upon contextual backing for their arguments created valid syllogisms, and others did not. The argument structures offered by each of these groups' participants who used primarily contextual backing are described in detail in this section.

The four participants who backed their arguments by reasoning that a basketball player's ability to make shots increases with the number of shots taken included context-related considerations such as the positive impacts of practice, muscle memory, and having less pressure if there are more shot opportunities. S57, for example, wrote:

> Completing at most 38.5% of shots in 13 attempts is more likely, as the more shots that he takes, the more comfortable he will become shooting. So, if he took 130 shots he would definitely be better than 38.5%. The player will get better with more shot attempts.

If one accepts the major premise that more practice leads to increased comfort and subsequently greater ability to convert shots, then the syllogism underlying responses like those given by S57 produces a valid argument for choosing the smaller sample. Such arguments can be concisely represented by removing the octagon from Figure 5 and changing the text in the backing (e.g., effects of practice, muscle memory, etc.) and other nodes to match those in the given response. S57's response was also notable in that "more likely" (which was in the problem statement) was used as an appropriate qualifier near the start of the response, but then the deterministic words "will" and "definitely" appeared in the latter part where qualifiers are still needed. So, in Figure 6, which represents S57's response, we introduce the convention of using a rectangular call-out node to indicate when deterministic words are used when qualifiers are needed instead.

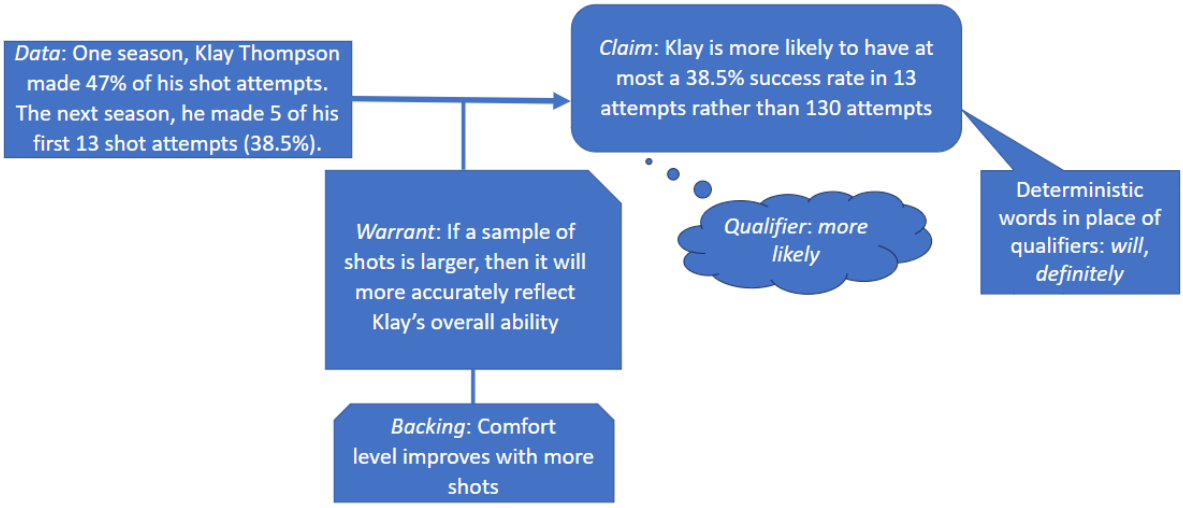

Figure 6. Representation of S57s argument structure with deterministic language callout

Notably, two of the four participants who provided contextual backing for more shots increasing a player's ability selected the larger sample (S18 and S61). S18, for instance, justified choosing the larger sample of shots by writing, "Having more shot attempts his nerves may be easier allowing him to make more shots." In such cases, even if the major premise were shown to be true, the syllogism underlying the response would not be valid. The essence of S18's argument, for example, was that the positive effect of taking many shots would lead to Klay's percentage being lower than normal; in such an argument, the major premise contradicts the conclusion. As noted in Section 4.2, the linguistic complexity of the problem may have contributed to such syllogistic difficulties. Such responses can be represented by leaving the octagon in Figure 5 in place and changing the text in the backing and qualifier nodes to match the contextual reasons and qualifiers in the response.

Five participants drew upon past observations of differences in game–to–game performance to back their arguments (S15, S40, S49, S63, S66). In these cases, the backing did not contradict the claim, but it also was not apparent how it would lead to movement from data to claim. S15, for instance, justified choosing the larger sample of shots by writing, "The second option because in some games he will shoot less than 47% and some games he will shoot over 47%. He will not consistently shoot 47%." Although this was a legitimate contextual observation about game–to–game variability, it was not apparent how it justified choosing 130 shots rather than 13. Others using game–to–game performance as backing chose the smaller sample of shots rather than the larger sample. These responses at times contained contextual vocabulary such as the presence of "bad games" (S63) or "off-nights" (S40) to characterize game–to–game variability, but such arguments also did not explain why at most 38.5% converted would be more likely in 13 shots rather than 130. The idea of variability, which is fundamental to the eLLN, was evident in such responses, but an intuitive expression of the eLLN (Brown, 2019) that would back a probabilistic claim about the relationship between sample size and population was not present. So, the contextual backing provided was essentially neutral in its potential to support movement from data to claim.

Four other participants (S14, S22, S29, and S55) offered neutral contextual backing by commenting on Klay being a skilled shooter. S14, for example, wrote, "The first option is more likely. He has a very good shot percentage. If he shoots 13 times he will shoot at least 38.5%. The chances are high that he gets better than that." In some of these cases, participants may have had eLLN intuitions in mind as they reasoned about the problem, but they did not explicitly state them. S22, for example, wrote, "#1 is more likely. Klay Thompson is a very skilled shooter and for him, making 5 out of 13 is much easier and more likely than 50 out of 130." S22 did not go on to state why 5 out of 13 was more likely than 50 out of 130, but the response is potentially compatible with eLLN reasoning. If pressed further, it is possible that participants like S22 may have expressed a version of the eLLN or contextual principles as backing for their arguments, but as stated, their contextual observations about Klay's skill were neutral in supporting movement from data to claim. Arguments with neutral backing are represented in

Figure 7, which has dashed lines connecting backing to warrant and data to claim to indicate that the backing was not contradictory to the movement from data to claim (as in Figure 5) but also did not provide impetus for the movement, either. Note that Figure 7 can also be used to represent arguments with neutral mathematical backing (third row of Table 1; Section 4.3 responses) by changing its node text to match the content of a given response.

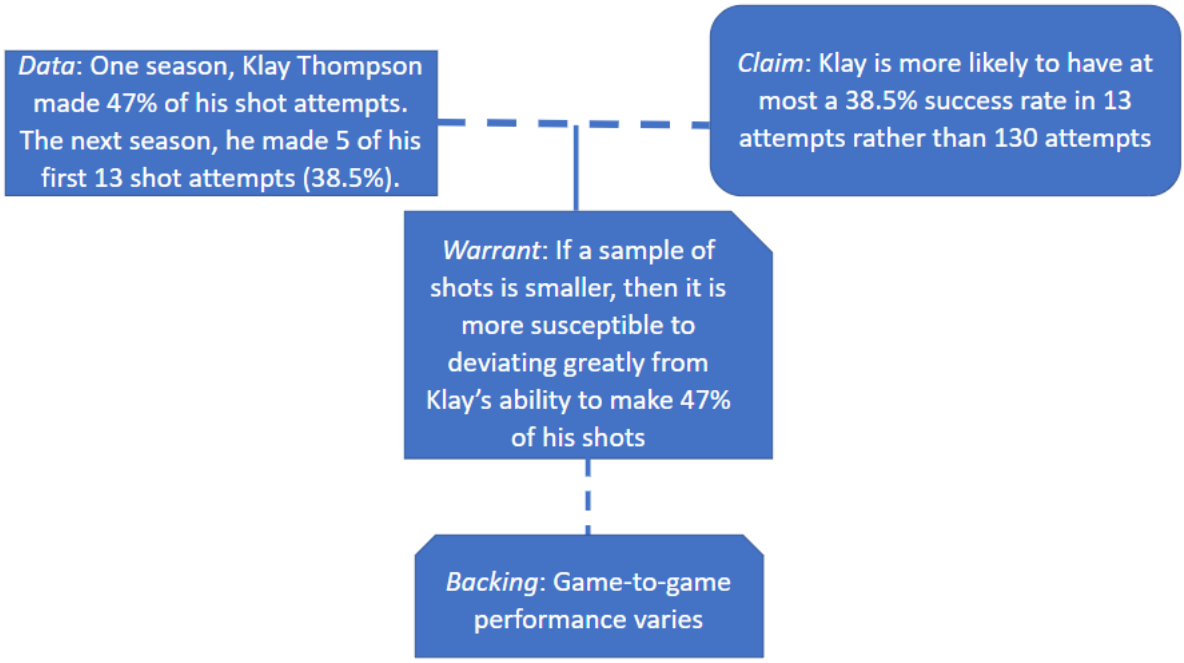

Figure 7. Diagram of an argument in which the backing provided is neutral in supporting movement from data to claim

In three cases (S17, S46, and S23), participants' neutral contextual backing included discussion of how players' conversion percentages are normally determined over the course of a season. S23, for example, justified choosing the larger sample by writing, "I chose the second one because I feel like usually NBA players make more than 13 shot attempts in a season. So they should take Thompson's average out of his total shots or larger amount than his first game." Although the contextual observations in such responses were accurate, they did not provide backing capable of supporting the move from data to claim; observing how shot percentages are normally determined, for instance, does not provide information about the amount of variability one would expect to see in small samples when compared to larger ones.

4.5. RESPONSES WITH NO EXPLICIT BACKING

In 12 responses (fifth row of Table 1), no explicit backing was readily apparent. Nine of these 12 argued that having more shots would increase Klay's success rate (S19, S31, S35, S37, S42, S45, S56, S59, and S65), and the other three said more shots would decrease Klay's success rate (S21, S32, and S33). S56, for example, chose the smaller sample, writing that "taking more shots can give you more opportunity to complete more." S33 chose the smaller sample as well but said, "taking 130 shots gives Thompson more opportunities to miss." In this category of response, participants did not explain why taking more shots would either lead to more makes or more misses. The 12 responses without explicit backing were equally split between choosing the smaller sample (S21, S32, S33, S45, S56, and S65) and the larger one (S19, S31, S35, S37, S42, and S59). If pressed, it is possible that these participants may have offered contextual backing such as practice, nerves, comfort, or muscle memory, as some others did (e.g., taking more shots allows you to make more *because you start to feel more comfortable shooting*). They may also have offered mathematical backing based on reasoning with absolute frequencies rather than proportional reasoning. Or, some may have offered an expression of the eLLN

as backing under additional probing. We thus characterized this set of responses as having no *explicit* backing, but the possibility that contextual, mathematical, or statistical backing may have been *implicit* in their reasoning cannot be ruled out.

Some of the arguments characterized as having no explicit backing contained valid syllogisms, and others did not. The argument structures for such responses can be represented by removing the backing and rebuttal nodes from Figure 4 and inserting text matching the warrant (e.g., if you take more shots, you have more success, or if you take more shots, you have more failure), claim (selecting either the larger or smaller sample), and qualifiers in the response in the appropriate nodes in Figure 4. An octagon can be inserted along the path from data to claim (as in Figure 5) to represent arguments containing invalid syllogisms.

4.6. RESPONSES BASED ON UNINTENDED INTERPRETATIONS OF THE PROBLEM

In 13 cases (sixth row of Table 1), context knowledge led participants to form unintended interpretations of the task. Arguments based on unintended interpretations can be modeled using the representational conventions we have discussed, although the substance of such arguments differed substantially from those in other categories. Three of these 13 participants (S30, S51, and S62) interpreted the problem to be asking if it was more likely to take 13 or 130 shots in one game. These three participants selected the sample of 13 shots because, as S51 explained, "no one takes 130 shots in one game." The problem itself did not state all 130 shots were to be taken in a single game, but these participants reformulated the problem in such a manner based on what they believed about its context.

Four others in this group of 13 participants (S1, S27, S47, and S52) reasoned that a basketball player's ability to make shots would decrease rather than increase when taking many shots in a short timespan, even though the problem did not state that the shots would be taken in a short period of time such as a single game. They explicitly mentioned fatigue as a factor in decreasing shooting ability and chose the smaller sample of shots. S1, for example, wrote: "I think option #1 (the smaller sample) is most likely to happen because Klay Thompson is a very good shooter but he would get fatigued if he tried to take 130 shots, therefore that would be a factor." In such responses, participants chose the intended sample, but they did so solely because it involved taking less shots in a short timespan, which they associated with less fatigue.

Four others (S39, S41, S43, and S48) of the group of 13 interpreted the problem to be one of predicting the likelihood of observing exactly 38.5% of shots completed in a single game. For instance, S39 chose the larger sample, writing, "Completing at most 38.5% of shots in 130 attempts would be more likely because Thompson would have more chances to receive 38.5%." In this case, it appeared that S39 also interpreted an "attempt" to refer to a game rather than an individual shot.

Finally, two in this group of 13 participants (S25 and S60) questioned the information provided in the problem based on their experience with the context. S25 questioned the accuracy of the data that Klay Thompson shot only 38.5% in a game, saying, "He is a professional athlete. Missing that much in only a few attempts doesn't seem right." S60 questioned the accuracy of the problem's use of 47% as an estimate of Klay's ability, writing, "Klay Thompson's previous percentage already shows his capability to go over the percentile of 47%." Such contextual considerations curtailed deeper engagement with the problem as it was written.

5. DISCUSSION

The research questions for the present study were: (1) What reasoning patterns are associated with participants' sample size choices for a hospital problem variant?; (2) How do participants use knowledge of problem context and the eLLN to reason about the variant? The columns of Table 2 summarize the observed participant reasoning patterns (research question 1) in terms of System 1 thinking (thinking fast), and System 2 thinking (thinking slow; Kahneman, 2011), and the rows summarize their use of context knowledge and the eLLN (research question 2).

Table 2. Summary of participants' reasoning patterns, use of context knowledge, and use of eLLN intuitions for a contextually provocative hospital problem variant

	System 1 (thinking fast)	System 2 (thinking slow)
Context	Ignored context or engaged it superficially	Used context knowledge to re-interpret the problem
	Used context knowledge without qualification	Used context knowledge with qualification
eLLN	Used the eLLN without qualification	Used one or more intuitive expressions of the eLLN with qualification

5.1. COMPARING PARTICIPANTS' SYSTEM 1 AND SYSTEM 2 THINKING

Some participant responses suggested automatic application of prior knowledge or intuitions characteristic of System 1 thinking. Those who offered responses without explicit backing (Section 4.5) may have taken such a cursory approach to the hospital problem variant used in the study. Participants who gave primarily mathematical backing (Section 4.3) seemed to perceive the variant as a school mathematics problem they had encountered in the past and quickly applied mathematical principles and operations to the numbers in the problem with little regard for context. Even some participants who based their responses mostly on the context of the problem (Section 4.4) at times applied their contextual intuitions without qualification where needed. S57, for example (Figure 6), used deterministic language to argue that a player's ability increases as they take more shots. Such responses lacked acknowledgement of possible limitations to the application of contextual ideas. Some participants also seemed to automatically apply eLLN intuitions to the situation without acknowledging potential problems with its application to the variant's context. For instance, S16 and some other participants whose responses are summarized in Section 4.1, used deterministic rather than qualified language in characterizing the eLLN's applicability to the variant. An ideal response, on the other hand, would involve engaging System 2 thinking to consider the extent to which the eLLN is applicable to the variant used in the present study.

System 2 thinking was apparent in some responses. Participants who used qualifiers in applying the eLLN to the variant appeared to have engaged System 2 thinking. For example, S28's response, and some others summarized in Section 4.1, used qualified language to suggest that the eLLN was likely to apply to the situation but did not frame its applicability in absolute terms. Such responses approximated normative statistical thinking more than other responses in our data set. Engaging System 2 does not, however, automatically produce arguments compatible with normative statistical discourse. S2, for example, used the eLLN in a qualified manner but had difficulty constructing a valid syllogism to support the argument (Section 4.2, Figure 5). Some using qualified context knowledge as backing (Section 4.4) failed to construct a syllogism in which the conclusion followed from the premises or constructed valid, but not necessarily true, arguments based on ideas like the effects of practice and comfort level. Participants who re-interpreted or questioned the statement of the problem using context knowledge (Section 4.6) also showed evidence of using System 2 thinking, as they compared their perception of the situation against that of the authors of the variant rather than quickly applying a previously constructed contextual, statistical, or mathematical script to produce a solution. System 2 thinking of this nature is necessary, but not sufficient, to produce a normative argument in response to the study's contextually provocative variant. Normative System 2 thinking about the variant requires complex, appropriately qualified coordination of context knowledge, statistical knowledge, and syllogism construction.

5.2. LIMITATIONS OF THE STUDY

Although the present study provides information about the arguments participants may use when approaching a contextually provocative (Madden, 2011) hospital problem variant, some limitations of

the study should be acknowledged. Results from the study are not statistically generalizable because we focused on a qualitative exploration of one group's responses to one variant. Participant and task characteristics both exert influence on hospital problem variant responses (Lem et al., 2011), so we would expect to observe different argument patterns for different populations and variants. There are also limitations associated with our use of writing prompts to collect data. We gathered data via written responses rather than interviews to encourage System 2 thinking (Kahneman, 2011; Menary, 2007), facilitate study of a relatively large sample, and make the activity fit naturally within the course activities for the physical education classes in which participants were enrolled. A limitation of the writing prompt approach is that some participants may have left argument aspects they had in mind unstated. For example, participants who did not back their warrants yet selected the smaller sample of shots as intended may have used eLLN intuitions without stating them. Given our methodology, the study results are best understood as representing participants' initial approaches to the hospital problem variant before further probing or instruction.

There are also limitations related to the qualitative data analysis procedures used for the study. The two authors of this study were specialists in different disciplinary fields. As a result, our individual analyses focused on noticing aspects of the data salient to each discipline rather than trying to independently produce the same characterizations of the data. Although this interdisciplinary approach ensured careful attention to both statistical and contextual aspects of participants' responses, those conducting similar studies in the future may benefit from adding intradisciplinary layers of analysis to the process. For example, a team of two or more statistics education researchers could do independent analyses of the data, compare their results, and compute measures of inter-rater reliability. Concurrently, a team of two or more context experts could do the same. After each intradisciplinary team reaches agreement on how the data could be characterized, the two teams could meet to compare their characterizations. These interdisciplinary meetings might lead to further rounds of intradisciplinary analysis, or they might be used as sites to collaboratively negotiate a final shared characterization of the data. Creating, testing, and refining such interdisciplinary qualitative data analysis procedures could provide valuable infrastructure for future studies of responses to contextually provocative items from various disciplines.

5.3. IMPLICATIONS FOR FUTURE RESEARCH

Despite the limitations of the study, the findings provoke careful consideration of how research involving hospital problem variants is conducted. In particular, the results challenge the conventional practice of relying exclusively or extensively upon the metric of the percent of participants offering the intended sample size choice (Lem, 2015; Weixler et al., 2019). This metric can produce false negatives and false positives regarding participants' use of eLLN intuitions. Of course, some false positives and negatives will always be generated by participants who choose a sample strictly by guessing, but the present study suggests some additional potential sources. As shown in our results, additional false negatives can come from participants who use eLLN intuitions as backing for an argument but have errant syllogisms underlying their reasoning. The complex linguistic structures of many hospital problem variants (Evans & DuSoir, 1977; Reagan, 1989) make such syllogistic reasoning a non-trivial matter whether one holds eLLN intuitions or not. Additionally, the present study illustrates that false positives about eLLN use can be generated by participants who choose the smaller sample using primarily contextual reasoning (e.g., fatigue, practice, muscle memory) and/or an unintended interpretation of the problem (e.g., believing 130 shots were to be taken in one game, not believing the data provided in the problem statement). Given the role of context knowledge in such reasoning patterns, the potential for similar false positives may be particularly high for contextually provocative variants.

In pointing out limitations of the metric of success rate choosing the intended sample, we do not mean to suggest that research using contextually provocative hospital problem variants should be abandoned. On the contrary, we learned that research with a contextually provocative variant can provide a useful window on students' integration of contextual and statistical knowledge. In many past studies, such insight about students' thinking has been gained by observing their activities during extended statistical investigations (e.g., Langrall et al., 2011; Pfannkuch, 2011; Shaughnessy & Pfannkuch, 2002). Contextually provocative hospital problem variants can yield some of the same types

of information (e.g., showing how contextual considerations can drive students' thinking) with the pragmatic advantage of taking less time to administer.

As research with contextually provocative hospital problem variants is carried out, however, it is important for researchers to expand their conceptualizations of participant success beyond just choosing the intended sample size. We observed various types of "success" beyond choosing the intended sample size in the present study. For example, some participants backed their arguments with multiple intuitive expressions of the eLLN (Brown, 2019), suggesting a deeper level of understanding than those who used just one expression. Some participants did not use the eLLN as backing, but instead gave contextual reasons, nonetheless constructed arguments with valid syllogisms. Testing the veracity of some of the contextual backing given in such arguments (e.g., effects of practice, muscle memory, etc.) could provide starting points for statistical investigations aimed at investigating the truth of the arguments. The use of qualifiers in arguments is another type of success not captured by the conventional success rate metric. The use of qualifiers in some arguments suggested that participants had given thought to conditions under which the claims in their arguments might not hold. Qualifying one's claims is characteristic of System 2 thinking that avoids the System 1 impulse to accept one's initial intuitions without questioning them. Engaging System 2 thinking in a hospital problem context is a significant accomplishment because the automatic activation of System 1 thinking has been conjectured as a primary cause for low success rates on such items (Kahneman, 2011). In sum, researchers can paint more comprehensive portraits of participants' reasoning with hospital problem variants by going beyond tabulating success in choosing the intended sample to investigate other dimensions such as depth of eLLN intuitions, syllogism construction, and qualification of arguments.

Along with expanding the number of dimensions of success considered for hospital problem variants, the present study suggests rethinking existing traditional characterizations of success on such problems. The use of the eLLN is usually associated with success on such problems, although, as noted earlier, we observed at least one instance where syllogistic reasoning prevented a participant with eLLN intuitions from offering a successful response. Moreover, in provocative, complex contexts, it is worth drawing a distinction between qualified and unqualified arguments using the eLLN. We observed both types of arguments in the present study. Qualified, rather than automatic, use of the eLLN is needed in contexts where assumptions warranting its application are questionable. In the basketball context, for instance, there have been several scholarly debates about the assumption of independence of shots (Bar-Eli et al., 2006). Qualified use of the eLLN in such a context, rather than unexamined application, is closer to what one would expect to see in professional statistical discourse. So, an important aspect of research with contextually provocative hospital problem variants is to examine how participants apply the eLLN and not just whether or not it is automatically applied in a manner reminiscent of System 1 thinking. Examining the qualifiers and rebuttals (Toulmin, 1958, 2003) in participants' hospital problem arguments, as modeled in the present study, provides a starting point for such investigations.

6. CONCLUSION

Although hospital problem variant research (Lem et al., 2011; Weixler et al., 2019) has been largely separated from emerging research on context knowledge in statistical reasoning (e.g., Langrall et al., 2011; Pfannkuch, 2011; Shaughnessy & Pfannkuch, 2002), the two strands of research can be complementary. The present study shows that choosing a hospital problem variant with a context that connects to the interests of a given participant group can provoke complex, interesting student arguments. The present study provides qualitative methods and representations that can be used to analyze the resultant argument structures, and it also suggests ways to reconceptualize and enrich research with hospital problem variants, especially those with provocative contexts. As this type of research continues, we can gain a progressively deeper understanding of the cognitive processes involved in coordinating statistical and contextual knowledge, which is fundamental to empirical enquiry (Wild & Pfannkuch, 1999).

REFERENCES

Arnold, P., & Franklin, C. (2021). What makes a good statistical question? *Journal of Statistics and Data Science Education, 29*(1), 122–130. https://doi.org/10.1080/26939169.2021.1877582

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of 'hot hand' research: Review and critique. *Psychology of Sport and Exercise*, *7*(6), 525–553. https://doi.org/10.1016/j.psychsport.2006.03.001

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K–12 guidelines for assessment and instruction in statistics education II (GAISE II): A framework for statistics and data science education*. American Statistical Association.

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance*, *24*(2), 245–257. https://doi.org/10.1016/0030-5073(79)90028-X

Ben-Zvi, D., & Aridor-Berger, K. (2016). Children's wonder how to wander between data and context. In D. Ben-Zvi & K. Makar. (Eds.), *The teaching and learning of statistics: International perspectives* (pp. 25–36). Springer. https://doi.org/10.1007/978-3-319-23470-0_3

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM Mathematics Education*, *44*(7), 913–925. https://doi.org/10.1007/s11858-012-0420-3

Brown, E. C. (2019). *Growing certain*: *Students' mechanistic reasoning about the empirical law of large numbers* (Publication No. 13895270) [Doctoral dissertation, University of Minnesota]. ProQuest LLC.

Chazan, D., Sela, H., & Herbst, P. (2012). Is the role of equations in the doing of word problems in school algebra changing? Initial indications from teacher study groups. *Cognition and Instruction, 30*(1), 1–38. https://doi.org/10.1080/07370008.2011.636593

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801–823. https://doi.org/10.1080/00029890.1997.11990723

Cornish, F., Gillespie, A., & Zittoun, T. (2013). Collaborative analysis of qualitative data. In U. Flick (Ed.), *The SAGE handbook of qualitative data analysis* (pp. 79–93). Sage.

delMas, R. C. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 79–95). Kluwer. https://doi.org/10.1007/1-4020-2278-6_4

Eldh, A. C., Årestedt, L., & Berterö, C. (2020). Quotations in qualitative studies: Reflections on constituents, custom, and purpose. *International Journal of Qualitative Methods, 19*. https://doi.org/10.1177/1609406920969268

Evans, J., & Dusoir, A. E. (1977). Proportionality and sample size as factors in intuitive statistical judgement. *Acta Psychologica*, *41*(2–3), 129–137. https://doi.org/10.1016/0001-6918(77)90030-0

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28*(1), 96–105. https://doi.org/10.5951/jresematheduc.28.1.0096

Gerbic, P., & Stacey, E. (2005). A purposive approach to content analysis: Designing analytical frameworks. *The Internet and Higher Education, 8*(1), 45–59. https://doi.org/10.1016/j.iheduc.2004.12.003

Gil, E., & Ben-Zvi, D. (2011). Explanations and context in the emergence of students' informal inferential reasoning. *Mathematical Thinking and Learning, 13*(1–2), 87–108. https://doi.org/10.1080/10986065.2011.538295

Groth, R. E., & Choi, Y. (2023). A method for assessing students' interpretations of contextualized data. *Educational Studies in Mathematics, 114*(1), 17–34. https://doi.org/10.1007/s10649-023-10234-z

Henriques, A., & Oliveira, H. (2016). Students' expressions of uncertainty in making informal inference when engaged in a statistical investigation using TinkerPlots. *Statistics Education Research Journal*, *15*(2), 62–80. https://doi.org/10.52041/serj.v15i2.241

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus, and Giroux.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Keith, W., & Beard, D. (2008). Toulmin's rhetorical logic: What's the warrant for warrants? *Philosophy & Rhetoric, 41*(1), 22–50. https://www.jstor.org/stable/25655298

Knipping, C., & Reid, D. (2015). Reconstructing argumentation structures: A perspective on proving processes in secondary mathematics classroom interactions. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education: Examples of methodology and methods* (pp. 75–101). Springer. https://doi.org/10.1007/978-94-017-9181-6_4

Langrall, C. W., Nisbet, S., Mooney, E., & Jansem, S. (2011). The role of context expertise when comparing data. *Mathematical Thinking and Learning, 13*(1–2), 47–67. https://doi.org/10.1080/10986065.2011.538620

Lem, S. (2015). The intuitiveness of the law of large numbers. *ZDM Mathematics Education, 47*(5), 783–792. https://doi.org/10.1007/s11858-015-0676-5

Lem, S., Van Dooren, W., Gillard, E., & Verschaffel, L. (2011). Sample size neglect problems: A critical analysis. *Studia Psychologica, 53*(2), 123–135.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage Publications.

Madden, S. (2011). Statistically, technologically, and contextually provocative tasks: Supporting teachers' informal inferential reasoning. *Mathematical Thinking and Learning, 13*(1–2), 109–131. https://doi.org/10.1080/10986065.2011.539078

Makar, K., & Confrey, J. (2004). Secondary teachers' reasoning about comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 353–373). Kluwer.

Masnick, A. M., Klahr, D., & Morris, B. J. (2007). Separating signal from noise: Children's understanding of error and variability in experimental outcomes. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 3–26). Lawrence Erlbaum Associates.

Maxara, C., & Biehler, R. (2010). Students' understanding and reasoning about sample size and the law of large numbers after a computer-intensive introductory course on stochastics. In C. Reading (Ed.), *Data and context in statistics education*: *Towards an evidence-based society*. *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July 2010), Ljubljana, Slovenia*. International Statistical Institute. https://iase-web.org/documents/papers/icots8/ICOTS8_3C2_MAXARA.pdf?1402524969

Menary, R. (2007). Writing as thinking. *Language Sciences, 29*(5), 621–632. https://doi.org/10.1016/j.langsci.2007.01.005

Miles, M. B., Huberman, A. M., & Saldaña, J. (2019). *Qualitative data analysis*: *A methods sourcebook* (4th ed.). Sage.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*.

Nilsson, P., Schindler, M., & Bakker, A. (2018). The nature and use of theories in statistics education. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 359–386). Springer. https://doi.org/10.1007/978-3-319-66195-7_11

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and Learning, 13*(1–2), 27–46. https://doi.org/10.1080/10986065.2011.538302

Reagan, R. (1989). Variations on a seminal demonstration of people's insensitivity to sample size. *Organizational Behavior and Human Decision Processes, 43*(1), 52–57. https://doi.org/10.1016/0749-5978(89)90057-5

Rossman, A., Chance, B., & Medina, E. (2006). Some important comparisons between statistics and mathematics, and why teachers should care. In G. F. Burrill & P. C. Elliot (Eds.), *Thinking and reasoning with data and chance: Sixty-eighth yearbook of the National Council of Teachers of Mathematics* (pp. 323–333). National Council of Teachers of Mathematics.

Rubel, L. (2009). Middle and high school students' thinking about effects of sample size: An in and out of school perspective. In S. L. Swars, D. W. Stinson & S. Lemons-Smith (Eds.), *Proceedings of the 31st Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 636–643). Georgia State University.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Lawrence Erlbaum Associates.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, *10*(1), 33–51. https://doi.org/10.1002/(SICI)1099-0771(199703)10:1<33::AID-BDM244>3.0.CO;2-6

Shaughnessy, J. M., & Pfannkuch, M. (2002). How faithful is Old Faithful? Statistical thinking: A story of variation and predication. *Mathematics Teacher*, *95*(4), 252–259. https://doi.org/10.5951/MT.95.4.0252

Sherlock-Shangraw, R. (2013). Creating inclusive youth sports environments with the Universal Design for learning. *Journal of Physical Education, Recreation, & Dance*, *84*(2), 40–46. https://doi.org/10.1080/07303084.2013.757191

Society of Health and Physical Educators. (2013). *National standards for K-12 physical education*.

Sommerhoff, D., Weixler, S., & Hamedinger, C. (2023). Sensitivity to sample size in the context of the empirical law of large numbers: Comparing the effectiveness of three approaches to support early secondary school students. *Journal for Didactics of Mathematics*, *44*(1), 233–267. https://doi.org/10.1007/s13138-022-00213-x

Stanovich, K.E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Tabor, J., & Franklin, C. (2019). *Statistical reasoning in sports* (2nd ed.). Bedford, Freeman, and Worth High School Publishers.

Tirosh, D., & Stavy, R. (2000). *How students (mis-)understand science and mathematics: Intuitive rules*. Teachers College Press.

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Toulmin, S. (2003). *The uses of argument* (updated edition). Cambridge University Press.

Warren, J. E. (2010). Taming the warrant in Toulmin's model of argument. *The English Journal*, *99*(6), 41–46. https://www.jstor.org/stable/20787665

Weixler, S., Sommerhoff, D., & Ufer, S. (2019). The empirical law of large numbers and the hospital problem: Systematic investigation of the impact of multiple task and person characteristics. *Educational Studies in Mathematics*, *100*(1), 61–83. https://doi.org/10.1007/s10649-018-9856-x

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, *47*(2), 289–312. https://doi.org/10.1016/0749-5978(90)90040-G

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

RANDALL E. GROTH
Salisbury University
Department of Secondary and Physical Education
1101 Camden Ave.
Salisbury, MD 21801
USA