Qualitative meta-analysis on the Hospital Task: Implications for Research

Jennifer Noll
Portland State University

Sashi Sharma
The University of Waikato

Journal of Statistics Education Volume 22, Number 2 (2014),
www.amstat.org/publications/jse/v22n2/noll.pdf

Key Words: Sample size; Variability; The law of large numbers; Statistical reasoning; Sampling distributions

Abstract

The "law of large numbers" indicates that as sample size increases, sample statistics become less variable and more closely estimate their corresponding population parameters. Different research studies investigating how people consider sample size when evaluating the reliability of a sample statistic have found a wide range of responses, from no consideration of sample size to over consideration of sample size. This paper provides a qualitative meta-analysis of studies that have used what we dub the "Hospital Task" for investigating peoples' thinking about the role of sample size in parameter estimation. This paper aims to investigate what the Hospital Task can tell us about how people make decisions under uncertainty and the usefulness of the task for developing models of students' statistical reasoning. To achieve these goals, we review the original task, synthesize results of other studies which have used some version of this task, provide a critique of the different versions of the task, discuss implications of the task for research, and provide insights and viewpoints from a small group of statisticians. The paper concludes with implications for further research.

1. Introduction

Stohl and Tarr (2002, p. 321) state that one of the most important aspects of formulating and evaluating inferences is understanding the *unpredictability* of random phenomenon in the shortrun but the *predictability* in the long-run trends in data (i.e., the law of large numbers). Hence, concepts of sample size and variability are important for students to consider when making inferences from a sampling distribution to a population and making connections between empirical and theoretical probability. According to Garfield and Ben-Zvi (2008), there are two major concerns of sampling: sampling representativeness and sampling variability, and both must "be understood and carefully balanced in order to understand statistical inference (p. 237)." Rubin, Bruce, and Tenny (1991) raise their concern that over dependence on sampling representativeness influences students to think that a sample tells everything there is to know about a population, while over reliance on sampling variability influences students to think a sample explains nothing important about a population. Other research on students' informal and formal inferential reasoning suggests that students have many difficulties in understanding statistical inference, which involves building a schema of many interrelated ideas such as representativeness, sampling variability and distribution (Saldanha and Thompson 2007; Shaughnessy 2007).

Statistics education researchers study student thinking about sample size and variability in order to build models of student development and better inform the teaching and learning of statistics. We hypothesize that as part of the process of studying student thinking, statistics education researchers develop statistics tasks they hope will produce useful data on student thinking. In fact, design research in mathematics education specifically focuses on designing mathematical tasks that serve to both influence student learning in particular directions and provide researchers with models of student learning (e.g., Gravemeijer 1998; Gravemeijer and Cobb 2006). Yet other research (e.g., psychology, cognitive science) may focus on how people process information or make decisions based on uncertainty rather than building models of student thinking (e.g., Kahneman and Tversky 1972). Different research goals mean that research tasks are developed for different purposes. Statistics educators' ability to construct models of student thinking, as well as theories of student learning, based on research are only as robust as the tasks developed by researchers. Tasks are often recycled, revised, refined and tested with different populations of students and for different purposes over a period of time. In statistics education, student responses on a particular task are studied to build frameworks of student reasoning. Typically, when different studies or multiple iterations of the same study use the same task and similar types of student responses are observed, then a research community may see the tasks as producing reliable sources of information about student reasoning. That is, frameworks (generated by researchers) that outline important aspects of student thinking and learning, based on observed student responses to a task, are seen as more reliable by the research community when multiple studies observe similar types of student responses (see examples in Shaughnessy 2007).

Thus, in order for the statistics education community to evolve, it is important, from time to time, to synthesize and critique prior research and the tasks used in the research. That is just what this paper aims to do. Specifically, the two questions guiding this paper are: (1) What do the results from the different "Hospital Task" studies tell us about how people reason and make decisions?; and (2) What can statistics educators glean from the results of the "Hospital Task" in terms of frameworks that explain students' statistical reasoning?

To answer the research questions posed, the paper takes the following approach: (1) summarize the results of Kahneman's and Tversky's (1972) work with the original version of the Hospital Task; and (2) synthesize and critique the body of research that has used the Hospital Task (or some version of it). We hypothesize that a thorough review of the Hospital Task and studies that have used it make a contribution to the literature because a thorough analysis of the strengths and weakness of the prior research facilitate more robust developments in future studies investigating student thinking about sample size and variability.

2. Summary of the Original Hospital Task

Kahneman's and Tversky's (1972) seminal work illustrated the undue confidence that people have in the reliability of small samples. In this study, they gave the Hospital Task (see Figure 1) as part of a larger survey to tertiary students in a quiz-like fashion. To answer the problem successfully, it is necessary to look beyond the proportions and appreciate the effect of sample size. Kahneman and Tversky were looking to see if students would choose option (b) (Figure 1), recognizing that smaller sample sizes are more likely to have greater variability.

Figure 1. The Original Version of the Hospital Task Kahneman and Tversky (1972, p. 443)

The Hospital Problem
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%, 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?

(a) The larger hospital
(b) The smaller hospital
(c) About the same (i.e., within 5% of each other)

* The original task had two versions, some with 'more than' and some with 'less than'. We focus our discussion on more than.

Kahneman and Tversky (1972) reported that fifty-six percent of the students judged the probability of obtaining more than 60% boys to be the same in the small and in the large hospital (option c)[1]. That is, the students answered that the two hospitals were equally likely to record an uneven proportion of births. According to Kahneman and Tversky, the representativeness heuristic underlies this misconception. People who rely on the representative heuristic tend to estimate the likelihood of events by neglecting the sample size or by placing undue confidence in the reliability of small samples. However, sampling theory suggests that the expected number of days on which more than 60% of the babies are boys is more likely to occur in a small hospital because a small sample size is more likely to digress from the true population percentage (50%). As the sample size increases, the sample statistic (here the proportion of boys born) is more

[1] We will refer to this option as the "equally likely" option. It is the option that has been dubbed as the "main misconception" option in much of the literature.

likely to approach the theoretical value for the entire population (Fischbein and Schnarch 1997). About 20% of the students made the appropriate choice (option (b)), whereas about 24% selected option (a)[2].

It should be noted that Kahneman's and Tversky's initial hypothesis was that people reason using particular heuristics, such as the representative heuristic, regardless of other information, such as sample size, that should influence their decisions. Applying the representative heuristic, "…event A is judged to be more probable than event B whenever A appears more representative" (p. 431). Kahneman and Tversky suspected the equally likely option would appeal to those who use the representative heuristic. They were not interested in studying *why* students make a particular choice of equally likely or large sample, or small sample, rather they were focused on the outcome of students' decisions. However, focusing only on the outcome of students' decisions is not as useful for researchers who want to build models of student reasoning.

3. Modifications to the Hospital Task: Concrete Sample Sizes

Since the Hospital Task first appeared in Kahneman's and Tversky's work, researchers (e.g., Evans and Dusoir 1977; Fishbein and Schnarch 1997; Watson and Moritz 2000), have criticized the original task (Figure 1) as ambiguous because sample sizes are not specified. These researchers modified the original task by adding in concrete sample sizes for the large and small hospitals. For example, Fischbein and Schnarch gave a version of the Hospital Task (Figure 2) that included more explanation of the problem, specifically providing the average number of babies born each day for the small and large hospitals. They also gave frequencies of boys born at the small hospital (more than 9 boys born during a day) and the large hospital (more than 27 boys born during a day), rather than just the percentages (more than 60% boys at each hospital). They gave this task to grades 5, 7, 9, and 11 students (20 students per grade level) and to 18 college students who were prospective teachers specializing in mathematics, none of whom had previously studied probability. Their sample represented a range of students with respect to socioeconomic level and cultural background. The task was administered to each group of students during regularly scheduled class time.

[2] Response (a) may imply that students believe that the variability of the sampling distribution increases with the sample size. In fact, this hypothesis was investigated in later studies using the Hospital Task and will be discussed later in the paper.

Figure 2. Fischbein's and Schnarch's (1997, p. 98) version of the Hospital Task

> In a certain town there are two hospitals, a small one in which there are, on the average, about 15 births a day and a big one in which there are, on the average about 45 births a day. The likelihood of giving birth to a boy is about 50%. (Nevertheless, there were days on which more than 50% of the babies born were boys, and there were days on which fewer than 50% of the babies born were boys.) In the small hospital a record has been kept during the year of the days in which the total number of boys was greater than 9, which represents more than 60% of the total births in the small hospital. In the big hospital, they have kept a record during the year of the days in which there were more than 27 boys born, which represents more than 60% of the births. In which of the two hospitals were there more such days?
>
> (a) In the big hospital there were more days recorded where more than 60% boys were born.
> (b) In the small hospital there were more days recorded where more than 60% boys were born.

The response the researchers were looking for is choice (b). They found mixed results and some new issues that did not surface in Kahneman's and Tversky's (1972) work. Table 1 provides a summary of the outcomes of Fischbein's and Schnarch's work.

Table 1. Summary results from Fischbein and Schnarch (1997, p. 98)

Response	Grade 5	Grade 7	Grade 9	Grade 11	College
Larger Hospital (option a)	20% (4/20)	35% (7/20)	5% (1/20)	10% (2/20)	0%
Smaller Hospital (option b)	0%	0%	5% (1/20)	0%	0%
Equally Likely (option c)	10% (2/20)	30% (6/20)	70% (14/20)	80% (16/20)	89% (16/18)
No Response	60% (12/20)	30% (6/20)	15% (3/20)	10% (2/20)	11% (2/18)
Other answers – not clear	10% (2/20)	5% (1/20)	5% (1/20)	0%	0%

Table 1 shows there were a high number of non-responses, particularly in the lower grades. Non-response rates for high school and college students were much lower. In addition, the younger students who did respond tended to choose the large hospital (option (a)); in contrast, the high school and college students in Fischbein's and Schnarch's study tended to select the equally likely option (c). In their study, only one student, a grade 9 student, selected the smaller hospital (option (b)). The results presented by Fischbein and Schnarch (1997) are similar to those of Kahneman and Tversky (1972) in that students at comparable grade levels (i.e., grade 11 students and college students) tended to select the equally likely option (option (c)). However, Fischbein and Schnarch were able to provide new information about responses over a wide range of grade levels. Specifically, lower grades tended to select the larger sample size.

Fischbein and Schnarch (1997) also generated new hypotheses with respect to the range of responses. They suggested that differences in students' responses throughout grade levels were a combination of instruction and intellectual development. The percentage of non-responses and option (a) responses can be expected to decline as students progress through school[3]. In addition, it is plausible that the percentage of students selecting the equally likely (option (c)) as opposed to the small sample size (option (b)) increases because mathematics instruction through the middle grades focuses heavily on proportions. Thus, it is likely that students see the problem as pertaining to their ability to recognize that the ratios are identical for the small and large samples, rather than about the law of large numbers (something they are unlikely to have studied at this point in their education).

4. Modifications to the Hospital Task: Perceptions of Representativeness

Other researchers (e.g., Bar-Hillel 1982; Evans and Dusoir 1977; Evans and Pollard 1982; Well, Polletsak, and Boyce 1990) have criticized the original version of the Hospital Task (Figure 1) for not being able to detect people's intuition about when a particular outcome is far enough away from the expected percentage to be considered non-representative. For example, is the 60% of boys born far enough away from the expected 50% to evoke people's perception of an unusual event? These researchers argued that Kahneman's and Tversky's (1972) task leaves open the question as to whether the students had no intuition for the representativeness of small samples or that they did not find 60% boys born to be non-representative. In other words, the original version of the task does not take into consideration how students may view a sample of 60% boys. Do students find 60% boys born to be unusual in anyway or is it too close to the expected 50%?

Evans' and Dusoir's (1977) version of the Hospital Task (see Figure 3) contained two modifications to the original task based on their hypothesis that the task wording was overly complex. First, they changed 'for a period of 1 year' to 'on one day.' Second, they changed 'more than 60% boys' to 'all boys.' The researchers predicted that these changes would clarify the problem so that students could more easily recognize the importance of sample size.

Eighty undergraduate students at a city polytechnic participated in Evans' and Dusoir's (1977) study and were randomly given one of four versions of the task. Twenty students were given the original version of the task, 20 students were given a version in which 'for a period of a year' was changed to 'one day,' 20 students were given a version in which 'more than 60% boys' was changed to 'all boys,' and 20 students were given a version that made both the previous changes ('one day' and 'all boys') (see Figure 3). In addition, in their version of the task, the equally likely option was omitted.

[3] Though not discussed in Fischbein and Schnarch, other researchers, such as Reaburn (2008) have hypothesized that young students may see option (a) as a natural option – when there are more items to choose from there is more variability. This issue will be addressed in later sections of the paper.

Figure 3. Evans' and Dusoir's (1977, p. 133) version of the Hospital Task

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day and in the smaller about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys born, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

Original Version: 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?

Modification 1: Which hospital do you think is more likely to find *on one day* that more than 60% of the babies born were boys?

Modification 2: For a period of 1 year, each hospital recorded the days on which *all* the babies born were boys. Which hospital do you think recorded more such days?

Both Modifications together: Which hospital do you think is more likely to find *on one day* that *all* the babies born were boys?

Eighty percent (48/60) of the students who had one or both of the 'clarified' versions of the task correctly chose the small hospital. However, only 55% (11 out of 20) of the students who had Kahneman's and Tversky's original version of the task gave the correct answer, the small hospital. Unfortunately, when Evans and Dusoir removed the equally likely option, it may have influenced student responses. Students may have felt they had to select either the large or small hospital and that equally likely was not an option. In addition, they do not report on the types of incorrect responses, so we do not know if any students responded that they could not determine an answer or that the chances are the same for the given percentage of boys born at either hospital. Thus, we do not have the ability to compare to the 'main misconception' as noted by Kahneman and Tversky.

Evans' and Dusoir's (1977) changes focused on a sample of babies that produced 100% boys and/or focused on the fact that the sample was based on one day rather than over an entire year, potentially allowing researchers the opportunity to get a sense of *when* students might find a particular result non-representative. Bar-Hillel (1982) made similar changes in content. He claimed that from a psychological standpoint, people may judge sample results as representative even if they deviate somewhat from the population parameters. That is, a person's judgment that a result of 60% is less likely in larger or smaller sample depends on whether the person perceives the result of 60% boys as non-representative. Bar-Hillel called this primary representativeness. Hence, 60% may not be perceived as non-representative (i.e., far enough away from 50%) and therefore does not evoke subjects' secondary sense of representativeness. Bar-Hillel tested this theory using variations of Kahneman's and Tversky's (1972) problem, including variations to the proportion of boys in the sample from 60% to 70%, and from 60% to 80%. Bar-Hillel gave the different versions of the task to undergraduate students. When Bar-Hillel used Kahneman's and Tversky's version of the task, the results of student responses were similar to what Kahneman and Tversky reported (students tendencies to select the equally likely option (c)). However, in the version of the task that reported the proportion of boys exceeded 70%, the modal response of students in his study became the smaller hospital (option (b)). Bar-Hillel concluded that once students perceive a sample result as non-representative, the sample size is more likely to play a normative role in people's evaluation. That is, students will conclude the smaller sample size is more compatible with non-representative results than a larger sample size.

The work of Bar-Hillel (1982) and of Evans and Dusoir (1977) show that a different wording of the task may make all the difference in student responses. At the very least, subsequent work with the Hospital Task has opened the question of just what researchers can conclude about student reasoning in the context of uncertainty from their responses to this task.

5. Modifications to the Hospital Task: Taking into Account Student Explanations

Finally, researchers (e.g., Noll 2007; Reaburn 2008; Sharma 1997; Watson and Moritz 2000) have criticized the original task because it did not provide insight into why students reason the way they do or how their reasoning might develop. Kahneman's and Tversky's goal was to investigate whether or not people intuitively use the law of large numbers when reasoning in probabilistic situations. They were not attempting to study why people reason the way they do or how such reasoning develops. Kahneman and Tversky (1972) constructed the Hospital Task as a forced multiple-choice format. This format allows students to have a correct response but does not provide researchers any information about how students understand the relation between sample size and variability. For instance, students who correctly responded that the small hospital is more likely to get a non-representative sample may have simply guessed.

Statistics educators are interested in additional details about why a student selected a particular option, how their choices and reasoning may change through schooling, and how teachers may use such knowledge to inform instruction. We argue that student explanations (for both correct and incorrect responses) enable researchers to build models of student reasoning to support the facilitation of correct conceptions. This section is devoted to research that adapted the Hospital Task in order to focus the investigation on the reasoning underlying a student's choice. This section is divided into two subsections. The first subsection focuses on synthesizing research about why students may select the larger sample or the equally likely response. The second subsection focuses on work that synthesizes conceptual frameworks for organizing our understanding of student thinking about the Hospital Task as well as how thinking on this task may progress throughout the grade levels.

5.1 Student Reasoning on Incorrect Choices

Watson and Moritz (2000) adapted the Hospital task by changing the context from a hospital to samples of students in a school setting (see Figure 4). Like Bar-Hillel (1982) and Evans and Dusoir (1977), they also added concrete sample sizes (50 versus 20) and changed the percentage to 80% boys rather than 60% in the original version of the task. Finally, they asked students to explain their thinking. Their analysis focused on student explanations rather than percentages of correct and incorrect responses. Their study contained a total of 62 students. There were 21 grade 3 students, 21 grade 6 students, and 20 grade 9 students from a variety of school regions, including suburban and rural schools. There were equal numbers of males and females. However, their version of the Hospital Task was given in an interview setting to a smaller subset (41 students) of their total sample (three grade 3 students, nineteen grade 6 and nineteen grade 9 students[4]).

Figure 4. Watson's and Moritz's (2000, p. 52) version of the Hospital Task

> Researchers went to two schools: One school in the centre of the city and one school in the country. Each school had about half girls and half boys.
> The researchers took a random sample from each school: 50 children from the city school, 20 children from the country school. One of these samples was unusual; it had more than 80% boys. Is it more likely to have come from:
>
> (a) The large sample of 50 from the city school, or
> (b) The small sample of 20 from the country school, or
> (c) Are both samples equally likely to have been the unusual sample?
> Please explain your answer.

Out of the 41 respondents, only 8 (19.5%) chose the small sample (option (b)), with only 6 of those being able to give adequate reasons. Eight out of 41 (19.5%) respondents selected the larger sample (option a). The most common student response in Watson's and Moritz's study was that of equally likely (61% or 25 out of 41 students), and the proportion of this response did not vary between the grade 6 and grade 9 students. Watson and Moritz did not specify a break down among grade level and response type.

On the one hand, because Watson and Moritz (2000) asked students to explain their reasoning, we have insight, not previously available, as to why students made a particular choice. For example, this study showed that students who selected the larger sample size tended to argue that there were more children to pick from, so there were more children to get the higher percentage of boys from (p. 66). Interestingly, a large percentage of the students in Fischbein's and Schnarch's (1997) study choose the larger sample size, as well as 25% of the students in Kahneman's and Tversky's (1972) study. Though the latter two studies did not collect data on students' reasoning, it is possible that the students in these studies had similar type reasoning. In addition, this type of response is reminiscent of additive reasoning, typical of young students on other types of probability problems. Students view the problem in terms of absolute frequencies where a larger sample size means more chance for variability (see Shaughnessy 2007). Thus, having this information on student reasoning allows us to compare with the results of other educational studies related to sample size and strengthens our knowledge of potential issues students may have when reasoning about sample size in relation to sampling variability.

On the other hand, there are significant differences in the context and language between the original version of the Hospital Task and the version used by Watson and Moritz (2000). The

[4] The small subset of students given the Hospital Task is due to the fact that it was the last of three questions on the interview protocol and not all students got to question 3 during the interview time, this is especially true of the grade 3 students.

context from hospital to school is probably not a major issue of concern in comparing the tasks. However, Watson and Moritz used new language in their version that may change how students interpret the problem. Specifically, they used the term sample and the term random. The addition of these new terms represents a significant change to the task and, potentially, to how students approach it. For example, on an earlier task in the same study, Watson and Moritz found that young students tended to think of a sample as a little bit of something (a little taste of something) and older students tended to think of sample as a little part of a whole. At least one student in their study selected the smaller sample size, suggesting that if you only sample a few rather than have the whole how would you know the average. This suggests that the student's interpretation of sample as a little part of the whole may have influenced his response. In addition, many of the younger students had not heard the term random before and older students had differing interpretations of what random means. Watson and Moritz reported that students who selected the equally likely option tended to give one of the following reasons: the process is random, or each school population from which the samples were taken had a 50% occurrence of each gender. We wonder, for the students who selected equally likely because of something they believed to be true about random events, how they would have responded to previous versions where random was not part of the phrasing? Thus, their version of the task raises the question of how the addition of the terms sample and random to the Hospital Task changed the nature of students' interpretations.

Reaburn's (2008) task (Figure 5) was adapted with respect to gender (focusing on the number of female births rather than male), average number of births at each hospital per day, percentage of female births on a given day (80% as opposed to lower percentages like 60 or 70 or not providing a specific number at all), and request for an explanation.

Figure 5. Reaburn's (2008, p. 415) version of the Hospital Task

> Half of all newborns are girls and half are boys. Hospital A records an average of 50 births per day. Hospital B records an average of 10 births per day. On a particular day, which hospital is more likely to record 80% or more of female births?
>
> (a) Hospital A
> (b) Hospital B
> (c) Both equally likely
>
> Please explain your reasoning.

Reaburn (2008) studied responses to the Hospital Problem as part of a wider study of students' intuitive statistical reasoning at an Australian university. The 26 participants were volunteers enrolled in a first year statistics unit. Of these 26 participants, one had studied mathematics at year 11, twenty had studied mathematics at year 12, one at Technical And Further Education (TAFE) and four at University. Students were given the Hospital Task during class time and student responses were analyzed for patterns in the explanations for their choices.

Three out of twenty-six (11.5%) students chose the incorrect answer (option (a)), ten out of twenty-six (38.5%) chose the incorrect option (equally likely option (c)), one student did not respond, and twelve out of twenty-six (46%) chose the correct option (option (b)). All the students who chose the large hospital (Hospital A) suggested that because there are more births to choose from, it is more likely that there will be more female births. Typical explanations included: “Because Hospital A has more births each day than Hospital B, it is likely that there will be more female births too.” This type of response is similar to student responses noted in prior research by Watson and Moritz (2000). This finding suggests that despite the fact that Watson and Moritz changed the task in a significant way, at least some of the student reasoning related to the selection of the larger hospital (or larger sample size) is consistent across versions.

Reaburn (2008) observed that two of the students who selected the equally likely option, (c), used reasoning that involved proportions or ratios. For example, “Percentage is independent of the total number of births, it is a proportion.” These explanations support the earlier hypothesis of Fischbein and Schnarch (1997) that students studying proportions may be inclined to see the task as relating to the fact that the proportions for each sample are equivalent. The other students who chose the equally likely option used reasoning that involved independence of probability for each individual outcome. For example, “The likelihood of gender is individual to the delivery not on the hospital and the number of babies the hospital delivers” (p. 418). This latter response is also similar to some of the student responses in Watson’s and Moritz’s (2000) study. Thus, there is some evidence from the two studies that, despite the wording of the task, some students may select the equally likely option because they are focused on the fact that we can expect 50% of births to be boys and 50% girls, just like in a school that contains 50% boys and 50% girls, we can expect samples taken from this group to maintain a similar proportion.

Finally, Reaburn (2008) observed that all the students who chose Hospital B indicated that it was more likely, or as some stated, “easier,” for Hospital B to deviate from the 50/50 ratio. For example, students made comments such as “it is more likely that 8 out of 10 births will be female than 40 out of 50, as it only requires that 3 additional births are female to be at 80%, instead of 15 additional female births when there are 50 total births” (p. 417). Again, Watson and Moritz (2000) observed similar responses from some of the students who selected the smaller sample size. Adding some evidence that at least a few students from the two studies may have interpreted the two different versions of the task in a similar way.

Reaburn (2008) also gave the Hospital Task to the same students after their statistics unit was complete. However, only 9 of the original 26 students took the post version. Of these nine, five had originally selected the equally likely option, but in the post version four of these five selected the smaller hospital. This suggests that instruction can help student intuitions. However, Reaburn’s sample is too small to draw any real conclusions. In addition, we do not know what happened in the instructional unit and if a version of the Hospital Task was used specifically in instruction.

5.2 Conceptual Frameworks of Student Reasoning

This section details studies that utilized some version of the Hospital Task, had students explain the reasons for their choices, and then built models of students’ thinking based on the categories that emerged in analysis. This section begins with a continuation of the Watson and Moritz (2000) study, describing the three-tiered framework of statistical literacy they developed.

Ensuing subsections outline additional frameworks from studies involving new populations: a non-Western culture, high school teachers, and graduate students, respectively.

Watson and Moritz (2000) asked students to explain the reasons for their choices and based on student responses classified student reasoning into a three-tiered framework of statistical literacy. Tier 1 is the development of a basic understanding of statistical terminology; Tier 2 is the development of an understanding of statistical language and concepts when they are embedded in the context of wider social discussion; and Tier 3 is the development of a questioning attitude one can assume when applying more sophisticated concepts to contradict claims made without proper statistical foundations (p. 48). They suggested that in order to progress from Tier 1 to Tier 2, students needed to develop an appreciation of variability and the concept of random selection, and in order to progress to Tier 3, students needed to develop an appreciation of the role of sample size and sample selection methods. Recall that their version of the Hospital Task included the language of sample and random. Take, for example, a student who suggested that either the small or large sample is equally likely to be the unusual one because the samples were drawn randomly from the population would be coded at Tier 2 because they have an appreciation of random selection as equaling the playing field and protecting against bias samples. However, such a response does not show an appreciation of the role of sample size in relation to sampling variability. Their framework provides curriculum planners a way to think about developing activities that can facilitate student development through the three tiers.

5.2.1 Non-Western Students' Perspectives

Sharma (1997), as part of a larger study, used a variation of the Hospital Task to explore Fijian high school students' understanding of sampling variation. Though her version of the task (see Figure 6) asked students to explain their answer, she changed the context of the problem to coins because she thought that the context of the hospital problem might confuse non-Western students. The secondary school selected for the research was an average high school in Fiji. The class consisted of 29 students aged 14 to 16 years of which 19 were girls and 10 were boys. Fourteen students were selected from this class and were representative of the larger group in terms of abilities and gender.

Figure 6. Sharma's version of the Hospital Task Sharma (1997, p. 161)

> Shelly is going to flip a coin 50 times and record the percentage of heads she gets. Her friend Anita is going to flip a coin 10 times and record the percentage of heads she gets.
> Which person is more likely to get 80% or more heads?
> Explain your answer.

Sharma (1997) created a three-category rubric to describe the results. The three categories in the rubric are: non-statistical, partial-statistical and statistical. Non-statistical responses were based on beliefs and experiences. Students with responses in the non-statistical category thought that the flipping of coins depends on luck or how one tosses the coin. For example, one student suggested that the speed at which the coin is thrown or the direction it swings will depend on whether or not it lands on heads or tails.

Student responses coded as partial-statistical applied rules inappropriately or based their reasoning on intuitions such as equiprobable and outcome approach (see Konold 1989). For example, students who appeared to use an outcome approach said that it was impossible to respond because it was "just a matter of chance." To be coded as statistical, student responses needed to include a discussion of the probability of getting a head or tail and what might happen in sequences of flips over the long run. For example, if a student suggested that it is easier for the ten flips to deviate from 50% to 80% because that only requires three additional tosses to be heads as opposed to fifteen additional tosses to be heads to get 80% heads with 50 tosses. None of the students' responses in Sharma's study were coded as statistical. Table 2 provides a break down for the numbers of students from each category as well as the numbers of correct responses[5].

Table 2. Student responses in Sharma's (1997) study.

	Non-Statistical	**Partial-Statistical**	**Statistical**
Correct Response (Anita)	0	0	0
Incorrect Response (Shelly)	4	2	0
Incorrect Response (equally likely or depends on luck)	2	5	0

It must be acknowledged that the limited use of probabilistic reasoning on this question may be a consequence of a lack of emphasis on variation in the Fiji curricular materials. Moreover, Sharma (1997) suggested these students were not used to explaining their thinking or perhaps had difficulty in explaining their reasoning in detail due to language difficulties. On the one hand, this population of students spanned a different culture in the sense that these were the first group of students from a non-Western educational background to be given the Hospital Task. Sharma suggested that their consideration of the context or lack there of may hold different consequences for how they respond to the problem than students who receive a Western education because students receiving a Western education are accustomed to solving problems that do not necessarily serve some purpose. Why would anyone want to know which person gets 80% or more heads? She argued that many students in Western cultures are used to being asked to solve mathematical tasks that lack purpose because they know how to operate within the school setting. On the other hand, Sharma removed the context for her study because she thought that the hospital context might confuse students from a non-Western culture. Thus, the responses in this study raise new questions. How do we use the Hospital Task in different cultural settings to get useful information on student thinking? Is there a weakness in the wording of Sharma's version of the task in that it is completely open-ended and does not focus the students to draw on other relevant knowledge? Students have to assume that heads and tails are equally likely. Perhaps including cues such as "fair" in the question would have aided in students' interpretation of this question.

[5] One student did not respond so the counts are out of 13.

5.2.2 Pre-service High School Teachers' Perspectives

Watson (2000) gave a version of the Hospital Task (see Figure 7) to 33 pre-service teachers (PSTs) who were all in a post-graduate teaching program. There was wide variation in the mathematics background of these students: 23 had at least studied mathematics up to the second year as part of their previous university courses (one of these was on leave from a PhD enrollment in mathematics), and 10 had a less extensive mathematics background. Unlike previous studies, participants were given the task to work on overnight but asked to complete it on their own. Watson's goal was to add to information provided from previous studies by sharing strategies of adults, specifically those planning to become high school mathematics teachers. She wanted to investigate how intuition versus mathematical calculations played into how students solved the problem and to see if there were gender differences in solution strategies.

Figure 7. Watson's (2000) version of the Hospital Task (p. 121)

> A town has two hospitals. On the average, there are 45 babies delivered each day in the larger hospital. The smaller hospital has about 15 births each day. Fifty percent of all babies born in the town are boys. In one year each hospital recorded those days in which the number of boys born was 60% or more of the total deliveries for that day in that hospital.
>
> Do you think that it's more likely that the larger hospital recorded more such days, that the smaller hospital did, or that the two recorded roughly the same number of such days?

Student responses were collected and Watson used a clustering technique (as in Miles and Huberman 1994) to analyze student reasoning. Students were divided into those who used intuitive reasoning only, mathematical reasoning only, or a combination of the two. Watson's results are summarized in Table 3. Intuition was defined as Fischbein and Schnarch (1997) did – "cognition that appears subjectively self-evident…" (p. 96). For example, the following response would be coded as intuitive, "I can't work this out mathematically, but intuitively I would say that the smaller hospital is more likely to have more variation from the average of 50% boys. …the more babies there are, the more likely the frequency of boys is to converge to 50% of births" (p. 125).

Mathematical reasoning was defined as using mathematical calculation. The mathematical reasoning code was subdivided into whether students utilized the binomial distribution or normal approximation to the binomial (formal), or more elementary mathematics, such as percentages (basic). For example, one student's basic mathematics argument was:

> Given that, for the larger hospital to record 60% or greater boys in a given day would require at least (27/30) 90% of the average number of boys born per day, as compared to only 30% (9/30) or more for the smaller hospital, one would suggest that it is more likely for the smaller hospital to record a greater number of such days.

Table 3. Watson (2000). Breakdown of student responses to the Hospital Task p. 125

Correctness of conclusion	**Type of Reasoning**		
	Intuition only	Mathematics only	Intuition and Mathematics
Correct (n= 18)	7	Formal: 2 Basic: 6	Formal: 3
Incorrect (n = 15)	8	Formal: 4	Basic: 3
Total	15	12	6

In terms of students successfully selecting the small hospital, Watson's results, 55% success rate, are higher then that of the university students in Kahneman's and Tversky's (1972) study (20%) and to the students surveyed by Fischbein and Schnarch (1997), (1%). This finding is not surprising given that the students in her study had stronger mathematics backgrounds. Watson did not observe any significant differences in responses of female versus male students on this task. Table 3 shows that the students who used mathematical arguments alone were more successful than those who used intuition alone. Yet, it is also apparent that mathematics alone was not a successful strategy for everyone. Some students made errors in their formal mathematical calculations but were unaware of their mistakes. Watson suggested that it was surprising not to see more students use intuition with formal mathematics. She argued that future teachers of mathematics need a certain amount of intuition for these kinds of problems so that they have a sense of when a mathematical solution makes sense. If the students who incorrectly applied the binomial distribution had some natural intuitions then perhaps they would have found the mistake in their calculations. Watson's (2000) work added to previous studies by exploring the thinking of students with formal mathematics backgrounds.

5.2.3 Graduate Student Perspectives

Noll (2007) used Watson's and Moritz's (2000) version of the task (Figure 5) to investigate graduate teaching assistants' (TAs') understanding of the connections between probability, sampling, statistical inference and the role of sample size in sampling variability. The task was given in survey form to 68 TAs from 18 different universities across the United States. The TAs in this study were enrolled in graduate mathematics (7 or 10.3%), statistics (54 or 79.4%), mathematics education (3 or 4.4%), or a related programs (4 or 5.9%) of study, and 97% had taken at least one graduate course in statistics. Noll conducted follow-up interviews with 5 TAs (volunteers from the same mathematics and statistics department). Noll reported that 55 of the 68 (80.9%) TAs who took the survey indicated that the small sample is more likely to be unusual. Yet, 12 TAs out of 68 (17.6%) responded that both samples were equally likely to be unusual, and one TA out of 68 (1.5%) indicated that the large sample was more likely to be unusual.

Noll developed a three category conceptual framework based on her analysis of the data: probability theory; sampling distributions; and, the outcome approach (as in Konold 1989). The probability theory approach consisted of two possible perspectives. The first perspective is connecting the context to a particular probability distribution (either applying a binomial or hypergeometric probability distribution). The second perspective is an application of the law of large numbers and suggests that as sample size increases sample variability tends to decrease.

Those TAs who indicated that the small sample is more likely to be unusual used either the binomial probability distribution argument (8.8% of the 55 TAs) or the law of large numbers (67.6% of the 55 TAs). Evidence for probability theory arguments were clear. For example, TAs gave explanations such as, “as sample size increases standard deviation decreases” or “using a binomial probability model I calculated the probabilities for 50 choose 40 versus 20 choose 16.” A sampling distribution approach was defined as an image of hypothetically repeating the experiment, collecting many sample statistics from samples of size 50 and from samples of size 20, and finally comparing the variability in the sampling distributions. Under this approach one would expect the sampling distribution for samples of size 50 to be less variable than for those of size 20. None of the TAs in this study used this approach.

The outcome approach indicates that the focus is predicting whether the particular sample of 20 is more or less likely to be unusual than the particular sample of 50. With this approach there is no focus on repeating the sampling process or thinking about what is likely to happen in the long run. Eleven of the 12 “equally likely” responses gave reasoning consistent with an outcome approach. These student explanations suggested that because both school populations had an equal ratio of boys and girls and the samples were randomly drawn, there was no reason to suspect that either sample was more likely to be unusual. The responses of these 11 TAs provides some indication that these TAs did not view the sample of size 50 and the sample of size 20 as one particular case of many possible sample cases. That is, these TAs may not have interpreted this problem as being related to the long-term relative frequencies. Rather, these TAs may have been answering the question: “For these two *particular* samples, one of size 20 and one of size 50, is one more likely to be unusual than the other?” In addition, this type of explanation is similar to what Watson and Moritz (2000) observed with some of the students in their study who selected equally likely. These TAs would be coded at Tier 2 by Watson’s and Moritz’s framework for statistical literacy

At the beginning of the interviews, Noll asked each of the TAs to explain how they interpreted the problem and to elaborate on the reasons for their choices. Two of the five TAs reasoned about this task from an outcome approach, and three of the five reasoned about this task from probability theory. The additional evidence of outcome approach responses from two of the five TAs she interviewed was compelling. Noll showed they did not have robust knowledge of the relationship between sample size and sampling variability. The TAs’ responses were based on the degree to which they believed they could predict the outcome for this one unique event. The participants argued that the samples were randomly selected and the ratio of boys to girls at each school was the same, so there would be no way to tell which sample was more likely to be unusual.

While the majority of TAs in the study responded in statistically appropriate ways to this task, the study is important for two reasons. First, it adds to the developmental frameworks from previous literature (e.g., Konold 1989; Watson and Moritz 2000) and, taken together, this research reveals a developmental progression from elementary school through advanced studies in statistics. For example, there is now literature that frames student reasoning in terms of statistical literacy skills (Watson’s and Moritz’s framework) and in terms of progressions of statistical thinking approaches (outcome approach - Konold, sampling distributions, probability theory - Noll). These frameworks suggest different ways to view student thinking, depending on the purposes of instruction and the grade level. In the early grades, for instance, we are focused on developing students' basic notions of statistical terminology and literacy skills. Teachers may need to classify where students fit within a hierarchy of reasoning. In contrast, as students progress in school, we may need to consider the implicit notions of probability that may underlie a particular view. Second, it is surprising that thirteen TAs (most of whom had numerous graduate statistics courses) did not respond in a normative way. The two TAs who used the outcome approach in the interviews were at the end of their master's programs in statistics. This suggests that the "equally likely" misconception may be quite robust despite significant amounts of statistics instruction. It also suggests that some teachers of statistics may perpetuate this misconception if they themselves have it.

6. Comparisons Across Versions of the Hospital Task

Just as with the original version of the Hospital Task, newer versions (e.g., Figures 2 & 3) have not gone without criticism from the research community. In particular, all versions of this task have been criticized in terms of what the task actually measures. Sedlmeier and Gigerenzer (1997) made a compelling argument that there is a significant difference between different versions of the task in terms of the type of distribution involved. They argued that in the original version the focus is on a sampling distribution because the task asks about a *proportion over the course of the year on multiple days*. In contrast, they suggested Evans' and Dusoir's (1977) version (Figure 2) focuses on the distribution of a sample (Sedlmeier and Gigerenzer used the language frequency distribution) because the problem asks about the sex of babies on a *particular day*. For ease of comparisons, Table 4 provides summary data from all of the studies reviewed in this paper.

Sedlmeier and Gigerenzer (1997) quoted Jacob Bernoulli, "even the stupidest man knows by some instinct of nature *per se* and by no previous instruction that the greater the number of confirming observations, the surer the conjecture" (p. 33). However, Kahneman's and Tversky's (1972) study and some of the other studies outlined here appear to contradict Bernoulli's statement. Sedlmeier and Gigerenzer constructed a hypothesis for why so many of the studies using the Hospital Task gave results that contradict Bernoulli's claim. Sedlmeier and Gigerenzer argued that people in general have natural intuitions when it comes to distributions of single samples and recognize that larger samples are less variable. They referred to this as intuition for the *empirical law of large numbers*. In their study, they categorized versions of the Hospital Task that focused on sampling distribution problems and those that focused on distributions of a single sample (i.e., frequency distributions), and concluded when the task was phrased in a way that focused on distributions of single samples, participants tended to perform much better (i.e., used natural intuitions that one large sample gives better information then one small sample). They argued that there are few instances in ordinary life where one would have a reason for collecting many samples and looking at the distribution of sample means from those samples. However, there are many instances where one would find reason for collecting a sample, recognizing that the larger the sample collected the better the prediction. Saldanha and Thompson (2003; 2007) claim that students, when asked to make judgments about outcomes of random sampling, tend to focus on individual samples and statistical summaries of them instead of how collections of sample statistics are distributed. In addition, they observed students' preferences for aggregating repeated samples into one large sample rather than looking at a distribution of sample means from repeated samples. This is consistent with Sedlmeier's and Gigerenzer's (1997) argument that really only statisticians have developed the concepts of using sampling distributions to study problems.

Gigerenzer (1996) also criticized Kahneman's and Tversky's (1972) work for focusing on only one interpretation of probability, that of a frequentist perspective and what should happen over the long term. Gigerenzer suggested that people may have natural tendencies to interpret probability situations with a singular perspective rather than a long-term frequency perspective. A singular perspective is characterized by a focus on the causal system that produced the particular outcome and by an assessment of likelihood based on "the propensities of the particular case at hand." In contrast, a distributional perspective relates the case at hand to a sampling schema and views an individual case as "an instance of a class of similar cases, for which relative frequencies of outcomes are known or can be estimated" (p. 512). Konold (1991) found strong empirical support for this conjecture. He provided compelling evidence that people, when asked questions about probability, think they are being asked to predict with certainty the outcome of an individual trial of an experiment. Konold refers to this perspective as the outcome approach. On the one hand, Falk and Konold (1992) point out that the formal probability is mostly concerned with deriving measures of uncertainty, or answering the question: *How often will event A occur in the long run?* On the other hand, what most people want is to predict what will occur in a single instance to answer the question: *Will A occur or not?* Thus, the goal in dealing with uncertainty is to predict the outcome of a single next trial rather than to estimate what is likely to occur after a series of events.

While the arguments made by Sedlmeier and Gigerenzer (1997) that characterize the different versions of the task from a frequentist vs. subjective perspective are compelling, we do not find them entirely convincing for explaining why results were better in versions of the task focused on single samples. First, many of these studies did not use similar populations of students (e.g., university students and/or spanned different countries). Student responses across age and culture contain significant sources of variability, such as normal mathematical maturity as well as type of curriculum materials. Second, the sample sizes used in the studies reported here were all relatively small and none of the studies incorporated a randomized control study of different versions of the task. Thus, any broad conclusions from these studies should be tempered in light of this knowledge. Third, sampling distribution versions of the tasks appear to be much more complicated linguistically because phrasing includes language about "the number of days on which the average…" as compared to a question about the average of a single sample. While Sedlmeier and Gigerenzer argued that it is not the complex wording of the task that makes it less understandable to students, rather the distinction between the sampling distribution and frequency distribution. We disagree. The phrasing required to indicate that the problem is about averages from many samples is inherently more complicated then a simple statement about the collection of one sample. Finally, as has already been mentioned, many of these studies did not ask students how they interpreted the problems or to explain their reasoning. Thus, we do not have robust data on how students interpret sampling distribution and frequency distribution problems.

Table 4. Synthesizing versions of the Hospital Task.

Study	Research Goal	Task Modification	Population	Overall Student Choices (in %)	Primary Conclusions
Kahenman and Tversky (1972)	Investigate people's heuristics for making judgments under uncertainty	Original task: Over the course of a year which hospital (small or large or equally likely) will record more days on which more than 60% of births are boys	Undergraduate Students (n=50)	Equally Likely Choice (Main 'Misconception'): 56% Large Sample Size: 24% Small Sample Size: 20%	Students tend to select equally likely. They conclude students do not consider sample size when considering sample representativeness.
Evans and Dusoir (1977)	Test four different versions of the task. Investigate whether a particular phrasing or percentage would lead to different results.	Wording modifications – 'on one day' vs. 'over year'; and 'all boys' vs. 'more than 60% boys'. Also removed equally likely option	Undergraduate Students (n=80)	Equally Likely Choice (Main 'Misconception') NOT AN OPTION Small Sample Size, using original version: 55% Small Sample Size, using one or more modifications: 80% Large Sample Size, using original version: 45% Large Sample Size, using one or modifications: 20%	Kahneman's and Tversky's version yielded similar results. Changes in language yielded more students selecting the smaller sample as non-representative
Fischbein and Schnarch (1997)	Investigate the evolution of students' statistical intuitions over grade levels.	Wording modifications – provided concrete numbers of boys for small and large hospitals	Grades 5, 7, 9, 11 and undergraduate students (n=98)	Equally Likely Choice (Main 'Misconception'): Middle School – 20% High School – 75% Undergraduate – 89% Large Sample Size: Middle School – 27.5% High School – 7.5% Undergraduate – 0% Small Sample Size: Middle School – 0% High School – 2.5% Undergraduate – 0% No Response or Other: Middle School – 52.5% High School - 15% Undergraduate 11%	Grades 5 & 7 largely did not respond or responded with equally likely or large sample options; Grades 9, 11 and college students tended to respond with equally likely option. As students get older tend to focus on equivalent ratios, perhaps a byproduct of instruction on ratio and proportion.
Bar-Hillel (1982)	Investigate primary representativeness: How extreme the percentage of boys born needs to be before students find a result unusual	Added in options for samples to contain 60%, 70%, 80% or all boys	Undergraduate Students (n=136)	Equally Likely Choice (Main 'Misconception'): For over 60% boys: 60% For over 70% boys: 32% For over 80% boys: 32% For all boys: 27% Large Sample Size: For over 60% boys: 20% For over 70% boys: 25% For over 80% boys: 26% For all boys: 19% Small Sample Size:	When proportion of boys born becomes 80% students tended to suggest those results came from the smaller sample. Otherwise similar to that of Kahneman and Tversky

				For over 60% boys: 20% For over 70% boys: 43% For over 80% boys: 42% For all boys: 54%	
Watson and Moritz (2000)	Investigate sampling tasks with young students and construct models of their thinking.	Provided concrete sample sizes and asked students to explain their thinking	Grades 3, 6, 9 (n=41)	Equally Likely Choice (Main 'Misconception'): 61% Large Sample Size: 19.5% Small Sample Size: 19.5%	Younger students reason additively, assuming that a larger sample size holds more opportunity for an unusual sample. As students get older tend to focus on equivalent ratios. Task was changed with new language (random & sample)
Reaburn (2008)	To investigate student reasoning behind the choices they made. To look for changes after a unit on statistics.	Provided concrete sample sizes; changed percentage to 80%; asked students to explain their thinking	Undergraduate students (n=26)	Equally Likely Choice (Main 'Misconception'): 38.5% Large Sample Size: 11.5% Small Sample Size: 46%	Many students tended to suggest the equally likely option. There was improvement on student choices after the intervention (4 out of 5 who had selected the equally likely option on the pre switched to the small sample size on the post).
Watson (2000)	Investigate sampling tasks with pre-service teachers and construct models of their thinking.	Provided concrete sample sizes and asked students to explain their thinking	Preservice high school teachers (n=33)	Equally Likely Choice (Main 'Misconception'): 30% Large Sample Size: 10% Small Sample Size: 54% Not Clear: 6%	About half correct responses, about half selected equally likely.
Sharma (1997)	Investigate sampling tasks with new population of students. Create frameworks of student thinking.	Task modified to context of coins; students asked to explain their thinking	Grades 9, 10, 11 (n=14)	Equally Likely Choice (Main 'Misconception'): 69% Large Sample Size: 31% Small Sample Size: 0%	Most of students had non-statistical responses to the problem. Predominately thought the answer could go either way because it was all based on luck. Coin context different from original version. Non-western population.
Noll (2007)	Investigate teachers of statistics thinking about sampling tasks. Create frameworks for their thinking.	Provided concrete sample sizes and asked students to explain their thinking	Graduate statistics students (n=68)	Equally Likely Choice (Main 'Misconception'): 17.6% Large Sample Size: 1.5% Small Sample Size: 80.9%	Most students saw small samples as more variable (using frequentist approach or a theoretical probability argument). Though some students appeared to use the outcome approach and conclude equally likely based on just two samples.
Sedlmeier and Gigerenzer (1997)	Test people's intuitions for empirical law of large numbers.	No modifications. Review other studies	No new population. They review other studies	Not applicable. No new population.	They reviewed studies that use frequency and sampling distribution versions of the task. They analyzed results showing students perform better on frequency distribution version. Suggested this is evidence of empirical law of large numbers.

In order to bring in additional viewpoints on what the different versions of the Hospital Task measure, we asked four statistician colleagues to share their thinking about different versions of the Hospital Task. They each suggested that the task was designed to assess a student's knowledge of variability. In addition, two also explained that the task could be used to assess students' knowledge of the law of large numbers as well as expected value. Interestingly, they all felt that different versions of the task measured the same idea and would produce similar responses in students. These four statisticians did not raise issues with frequency distribution versus sampling distribution versions of the task. It could be that for expert statisticians, sampling distribution concepts are completely intuitive after years of practicing statistics and, thus, they do not see why students might struggle with ideas of sampling distributions more than ideas of frequency distributions.

We also asked these statisticians whether they thought the task was useful in supporting students' development of important statistical ideas of variability. Two stated that it was useful for helping students to understand the role of sample size in sampling variability. One statistician suggested that the usefulness of the task depended on how the task was introduced in the classroom. She suggested that while it is important for students to learn about variability, and that larger sample sizes produce more consistent results, this type of question, as written, "seems counter intuitive" unless tied to decision making contexts and how this knowledge can help one make reasonable decisions under uncertainty. Finally, one statistician did not believe the task was useful because he did not believe it was a good predictor of students' statistical intuition, rather he believed the task was more of a "trick" problem. He suggested that the wording leads students to select the equally likely response unless they were trained with tasks like the Hospital Task during instruction. Some of the opinions of these statisticians overlap with critiques of how this task has been used in the research literature. We still have little information about how students from these studies interpreted the question, nor how they view probability (e.g., outcome approach or frequentist approach).

7. Discussion and Conclusions

In this paper, we described the original version of the Hospital Task (recall Figure 1). This task was given to undergraduate students. The main conclusion was that students tended to select the equally likely option (the non-representative sample is equally likely to come from a small or large sample), suggesting people do not recognize the role of sample size in sample variability. There were three primary criticisms of the original task, which led to subsequent modifications and further research studies. First, the task was deemed overly complex in part because it did not provide students with concrete sample sizes for the large and small sample. Second, the task only provided 60% for the proportion of the boys born, missing an opportunity to investigate whether or not students found this percentage far enough away from the expected 50% to consider it non-representative. Third, the task was a forced-choice problem with no opportunity to query students as to the reason for their choices; thus, a missed opportunity to understand student thinking. In addition, there was some criticism across studies. Sedlmeier and Gigerenzer (1997) criticized research that used versions of the task that were constructed using sampling distribution language (i.e., average boys born over a year). They also criticized studies that employed multiple versions of the task (versions containing sampling distribution language and frequency distribution language), if the researchers did not acknowledge that the complexity was based in people's lack of natural intuition for sampling distribution as opposed to complexity in the wording of the task.

Returning to the first research question posed at the beginning of this paper: What do the studies utilizing the Hospital Task tell us about how students reason about uncertainty? Looking for common themes across Table 4 provides some indication that the equally likely option is appealing to high school and college students with limited mathematics backgrounds. Many of the studies involving high school and college students found 50% or more of the students selecting the equally likely option. The percentage of students selecting equally likely did appear to drop when the task indicated larger percentages of boys born (70% or 80%) or when the task talked about babies on one day versus over the year. In addition, the data summarized in Table 4 shows that younger students may have tendencies to select the larger sample size. Yet, we find that this research is still in its infancy and there are numerous limitations. The original study was focused on telling us what people's natural tendencies regarding sample size and sampling variability were. The task was not designed to give us insight into how students think statistically, nor how they interpret the task. Subsequent studies provided further insights into how students might interpret this task (or variations of it) as well as the reasoning behind their selections. However, subsequent studies involved relatively small samples of students where researchers had convenient access to the populations. Thus, results from these studies are tentative.

However, these limitations provide an opportunity for future research. The work that has been done provides a road map for future studies. For example, our review of studies involving the Hospital Task suggest that (1) the phrasing of the task; (2) task context; (3) statistical implications of the task (sampling distribution versus frequency distribution); (4) having multiple versions and multiple tasks to test task efficacy; (5) participants' interpretations of task; (6) mathematical background of participants; and, (7) culture of participants are all vital components if we are to design tasks that the statistics education research community can use to construct models of student thinking and development and, subsequently, use such models to design quality statistics curriculum. Particular directions include scaling up studies with larger populations of students, cross cultural studies to further investigate possible differences there, randomized controlled trials to better compare and contrast different versions of the tasks, and more qualitative analysis into student reasoning with this task. Scaling up studies would allow researchers to begin to generalize findings to larger groups of students. Cross cultural studies involving the Hospital Task (as well as other forms of sampling tasks) may be useful for informing statistics educators of how students' statistics knowledge develops in coordination with their particular cultural contexts. Greer and Mukhopadhyay (2005) caution us that statistics has been introduced with minimal regard to historical and cultural contexts in spite of the fact that it is a cultural construction. Randomized controlled trials with different versions of the task would allow researchers deeper insight into how students interpret different versions of the task and may provide further insight into Sedlmeier's and Gigerenzer's (1997) hypothesis of people's intuitions for the empirical law of large numbers.

Finally, the collection of these findings into a synthesized whole is informative for the field with respect to the second question addressed in this paper: Can statistics educators use the Hospital Task to construct helpful conceptual frameworks about student thinking? Worthwhile among the findings are that students' reasoning changes over time. For example, there may be a natural development from a focus on total frequency - 'more data more opportunity for variation;' to a focus on proportions – 'same proportion means equally likely;' to a focus on the role of sample size – 'large samples tend to be more representative.' In addition, the research provides some categorization of reasoning about probabilistic settings: the outcome approach, the frequentist approach, and a formal probability approach. These findings give the field an initial framework for the kinds of reasoning students apply to this problem, including reasoning types at different grade levels (elementary school through graduate statistics students). The initial framework also provides a starting point for future investigations that could refine, revise and extend the current framework. Furthermore, the framework and results of the various studies synthesized here can be of value to teaching. For example, if we know that young students tend to start off reasoning about these kinds of problems based on the larger the sample the more variation we can see, then we can sequence instruction to build toward proportional reasoning skills or hands-on simulations where students can take small and large samples and observe the results. Likewise, if we know that students may hold differing interpretations of probability in different settings then we can address the underlying notions of probability more explicitly in instruction.

References

Bar-Hillel, M. (1982), "Studies of Representativeness," In D. Kahneman, P. Slovic, and Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, New York.

Ben-Zvi, D. (2004), "Reasoning About Variability in Comparing Distributions," *Statistics Education Research Journal, 3(2)*, 4-6.

Evans, J. St B. T., and Dusoir, A. E. (1977), "Proportionality and Sample Size as Factors in Intuitive Statistical Judgement," *Acta Psychologica*, *41*, 129–137.

Evans, J. St B. T., and Pollard, P. (1982), "Statistical Judgement: A Further Test of the Representatives Construct," *Acta Psychologica*, *51*, 91–103.

Falk, R., and Konold, C. (1992), "The Psychology of Learning Probability," In F. Sheldon and G. Sheldon (eds.), *Statistics for the Twenty-First Century*, pp. 151-164, Washington: Mathematical Association of America.

Fischbein, E., Nello, M. S., and Marino, M. S. (1991), "Factors Affecting Probabilistic Judgements in Children and Adolescents," *Educational Studies in Mathematics, 22*, 523-549.

Fischbein, E., and Schnarch, D. (1997), "The Evolution with Age of Probabilistic, Intuitively Based Misconceptions," *Journal for Research in Mathematics Education*, 28, 96-105.

Garfield, J.B., and Ben-Zvi, D. (2008), "Preparing School Teachers to Develop Students Statistical Reasoning," In C. Batanero, G. Burrill, C. Reading & A. Rossman (eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics, Challenges for Teaching and Teacher Education, Proceedings of the ICMI Study 18 and 2008 IASE Roundtable Conference*

Gigerenzer, G. (1996), "On Narrow Norms and Vague Heuristics: A Rebuttal to Kahneman and Tversky," *Psychological Review*, *103*, 592-6.

Gravemeijer, K. (1998), “Developmental Research as a Research Method. In *Mathematics Education as a Research Domain: A Search for Identity*, pp. 277–295, Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-94-011-5470-3_18

Gravemeijer, K., and Cobb, P. (2006), “Design research from a Learning Design Perspective,” *Educational Design Research*, 17–51.

Greer, B. (2001), "Understanding Probabilistic Thinking: The legacy of Efraim Fischbein," *Educational Studies in Mathematics, 45*, 15-53.

Greer, G., and Mukhopadhyay, S. (2005), "Teaching and Learning the Mathematization of Uncertainty: Historical, Cultural, Social and Political Contexts," In G. A. Jones (ed.), *Exploring Probability in School: Challenges for Teaching and Learning*, pp. 297-324, New York: Springer

Jones, G. A., Langrall, C. W., and Mooney, E. S. (2007), "Research in Probability: Responding to Classroom Realities," In F. K. Lester Jr (ed.), *Second Handbook of Research on Mathematics Teaching and Learning*, pp. 909-955, Reston: The National Council of Teachers of Mathematics

Kahneman, D., and Tversky, A. (1972), "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology*, 3, 430-454.

Kahneman, D., and Tversky, A. (1982), "Variants of Uncertainty," In D. Kahneman, P. Slovic, & A. Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*, pp. 509-521, New York: Cambridge University Press.

Konold, C. (1989), “An Outbreak of Belief in Independence?,” In C. Maher, G. Goldin & B. Davis (eds.), *The Proceedings of the Eleventh Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, Vol II (pp. 203-209). Rutgers, Rutgers University Press.

Konold, C. (1991), "Understanding Students’ Beliefs about Probability, "In E. von Glasersfeld (ed.), *Constructivism in Mathematics Education*, pp. 139–156, Dordrecht, The Netherlands: Kluwer.

Lubienski, S. (2002), "Research, Reform, and Equity in US Mathematics Education," *Mathematical Thinking and Learning*, 4(2/3), 103–125.

Miles, M. B., and Huberman, A. M. (1994), *Qualitative Data Analysis: An Expanded*

Sourcebook, (2nd ed.). Thousand Oaks, CA: Sage Publications.

Nasir, N. S., Hand, V., and Taylor, E. V. (2008), "Culture and Mathematics: Boundaries between "Cultural" and "Domain" Knowledge in the Mathematics Classroom and Beyond," *Review of Research in Education*, 32, 187-240.

Noll, J. (2007), Graduate Teaching Assistants' Statistical Knowledge for Teaching. Unpublished doctoral dissertation from Portland State University.

Noll, J. (2011), "Graduate Teaching Assistants' Statistical Content Knowledge of Sampling," *Statistics Education Research Journal*, 10(2), 27-47.

Reaburn, R. (2008), "The Hospital Problem Revisited. Tertiary Student's Perceptions of a Problem Involving the Binomial Distribution," In M. Goos, R. Brown, and K. Makar (eds.) *Proceedings of the 31st Annual Conference of the Mathematics Education Research Group of Australasia*, © MERGA Inc. 2008

Rubin, A., Bruce, B., and Tenney, Y. (1991)," Learning about Sampling: Trouble at the Core of Statistics," In D. Vere-Jones (ed.), *Proceedings of the Third International Conference on Teaching Statistics. Vol. 1. School and General Issues*, pp. 314-319, Voorburg, The Netherlands: International Statistical Institute.

Saldanha, L., and Thompson, P. (2003), Conceptions of Sample and their Relationship to Statistical Inference. *Educational Studies in Mathematics*, 51, 257-270.

Saldanha, L., and Thompson, P. (2007), "Exploring Connections Between Sampling Distributions and Statistical Inference: An Analysis of Students' Engagement and Thinking in the Context of Instruction Involving Repeated Sampling," *International Electronic Journal of Mathematics Education*, 2, (3), www.iejme.com

Sedlmeier, P. (1998), "The Distribution Matters: Two Types of Sample-size Tasks," *Journal of Behavioral Decision Making*, 11, 281-301.

Sedlmeier, P., and Gigerenzer, G. (1997), "Intuitions about Sample Size: The Empirical Law of Large Numbers," *Journal of Behavioral Decision Making*, 10, 33-51.

Sharma, S. (1997), "Statistical Ideas of High School Students: Some Findings from Fiji," Unpublished doctoral thesis. Waikato University, Hamilton, New Zealand.

Shaughnessy, J. M. (2007), "Research on Statistics Learning and Leasoning," In F. K. Lester (Jr ed.), *Second Handbook of Research on Mathematics Teaching and Learning*, pp 957-1009, Reston: The National Council of Teachers of Mathematics.

Stavy, R., and Tirosh, D. (2000), *How Students (Mis-)Understand Science and Mathematics Intuitive Rules*, New York: Teachers College Press.

Stohl, H., and Tarr, T. J. (2002), "Developing Notions of Inference using Probability Simulation Tools," *Journal of Mathematical Behavior*, 21, 319–337.

Tversky, A., and Kahneman, D. (1974), "Judgement Under Uncertainty: Heuristics and Biases," *Science*, 185, 1124-1131.

Watson, J. M. (2000). “Preservice Mathematics Teachers’ Understanding of Sampling: Intuition or Mathematics”, *Mathematics Teacher Education and Development*, 2, 121-135.

Watson, J. M., and Moritz, J. B. (2000), "Developing Concepts of sampling," *Journal for Research in Mathematics Education*, 31(1), 44-70.

Well, A. D., Pollatsek, A., and Boyce, S. J. (1990), "Understanding the Effects of Sample Size on the Variability of the Mean," *Organizational Behavior and Human Decision Processes*, 47, 289-312.

Jennifer Noll
Assistant Professor
Fariborz Maseeh Department of Mathematics and Statistics
Portland State University
503-725-3643
noll@pdx.edu

Sashi Sharma
Senior Lecturer
Department of Mathematics, Science and Technology Education
The University of Waikato
Hamilton
New Zealand
phone: +61 7 838 4500 ext 6298
email: sashi@waikato. ac.nz

Volume 22 (2014) | Archive | Index | Data Archive | Resources | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Guidelines for Readers/Data Users | Home Page | Contact JSE | ASA Publications