STUDENTS' INFORMAL HYPOTHESIS TESTING IN A PROBABILITY CONTEXT WITH CONCRETE RANDOM GENERATORS

PER NILSSON
Örebro University, Sweden
per.nilsson@oru.se

ABSTRACT

This study examines informal hypothesis testing in the context of drawing inferences of underlying probability distributions. Through a small-scale teaching experiment of three lessons, the study explores how fifth-grade students distinguish a non-uniform probability distribution from uniform probability distributions in a data-rich learning environment, and what role processes of data production play in their investigations. The study outlines aspects of students' informal understanding of hypothesis testing. It shows how students with no formal education can follow the logic that a small difference in samples can be the effect of randomness, while a large difference implies a real difference in the underlying process. The students distinguish the mode and the size of differences in frequencies as signals in data and used these signals to give data-based reasons in processes of informal hypothesis testing. The study also highlights the role of data production and points to a need for further research on the role of data production in an informal approach to the teaching and learning of statistical inference.

Keywords: *Statistics education research; Informal statistical inference; Informal hypothesis testing; Experimentation-based teaching; Probability distribution; Inferentialism*

1. INTRODUCTION

Drawing a statistical inference is the act of moving beyond the data at hand to draw a general conclusion about some wider universe, while taking into account that variation is everywhere and that the conclusions are therefore uncertain (Moore, 2004). Garfield and Ben-Zvi (2008) differentiate between two primary situations in statistical inference: parameter estimation and hypothesis testing. Parameter estimation aims at characterizing a population (e.g., shape, center) based on a sample of data. In hypothesis testing, we are not directly interested in describing the actual population or process from which the samples come; instead we make a guess about the value of the parameter and then we test this guess by looking at how likely an observed result would be if the guess were correct (Konold & Pollatsek, 2002). If this probability is very low, we have two scenarios: (i) the result was caused by chance and even though it had low probability the sample belongs to the population with our guessed parameter value. This implies that we cannot reject our guess about the population parameter. And, (ii) it is safer to infer that a result with such a low probability would not really have happened by chance and thus, the sample result does not belong to the population with our guessed parameter value why we reject the guess (Kula & Koçer, 2020). The focus of the present study is on *informal hypothesis testing* (IHT).

Given the importance of understanding inferential statistics (Zieffler et al., 2008), as well as the consistent difficulties people have with this type of reasoning (Dolor & Noll, 2015), it is recognized widely that the foundation of fundamental ideas of inferential statistics should be laid in the early years of schooling (Meletiou-Mavrotheris & Paparistodemou, 2015). Exploring and characterizing statistical inference using an informal approach has been advocated in earlier grades (Makar & Rubin, 2009), with students being engaged in situations by producing samples (Doerr, Delmas, & Makar, 2017; Meletiou-Mavrotheris & Paparistodemou, 2015) and using informal methods of making statistical inferences (Zieffler et al., 2008).

This study examines fifth-grade students' IHT in the context of drawing inferences of underlying probability distributions. The usefulness of engaging students in informal inference tasks underlying

Statistics Education Research Journal, 19(3), 53-73, http://iase-web.org/Publications.php?p=SERJ

probability distributions has received increasing attention (e.g., Pratt et al., 2008; Stohl Lee et al., 2010; Zieffler et al., 2008). In a probability context, hypothesis testing may entail inferring whether a sample comes from a uniform probability distribution or if it is more likely that it is the result of a non-uniform probability distribution. Through a small-scale teaching experiment, the study addresses the following research questions: How do students distinguish a non-uniform probability distribution from uniform probability distributions in a data-rich, experimentation-based learning environment, and what role do processes of data production play in their investigations?

2. THEORETICAL APPROACH

2.1. INFERENTIALISM

Inferentialism is a theory attracting increasing attention in statistics education (Nilsson et al., 2017). It is a semantic theory, putting inference at the core of human knowing (Brandom, 1994), and thus fits well with the idea of statistical inference at the heart of statistical knowing (Bakker & Derry, 2011). The theory builds on Sellars's philosophical idea that the meaning and understanding of concepts[1] and claims are to be understood in relation to how they are used in reasoning (Sellars et al., 1997). For instance, as reasonable beings, we have the capacity to understand that claiming "The probability is fifty percent" is compatible with "The probability is 0.5" but incompatible with "The probability is one-third."

To Brandom (1994), meaning making is essentially an inferential practice: concepts, and claims in general, acquire their meaning in relation to other concepts and claims. Consequently, and by necessity, inferentialism is resolutely holistic (Bakker & Derry, 2011; Bransen, 2002). Think of the meaning of the concept "probability", being articulated in the claim, "The probability of an outcome of six is 1/6 when rolling a die." Claiming this implies, among many things, knowing that the claim is based on a fair die, that the relative frequency of sixes stabilizes around 1/6 as we increase the number of trials, and that the probability of not having a six is 5/6. This example involves inferential relationships between many reasons related to the concept of probability, of which only a few have been made explicit here. The main point, however, is that since claims both serve as and stand in need of reasons, concepts involved have the contents they have in virtue of the role they play in reasoning (Brandom, 2000).

To Brandom (1994), knowing is socially enacted; rather than being articulated in the space in which thought moves, knowing is articulated in the space of a language game (Sellars et al., 1997; Wittgenstein, 1968) and, particularly, in the *game of giving and asking for reasons* (GoGAR) (Brandom, 1995, 2002). In other words, understanding a concept (e.g., "probability" in "the probability of six is 1/6 when rolling a die") entails understanding the conditions of its application in the GoGAR (Bakker & Derry, 2011; Bransen, 2002). It is "to have practical mastery over the inferences [the concept] is involved in—to know, in the practical sense of being able to distinguish, what follows from the applicability of a concept, and what it follows from" (Brandom, 2000, p. 48). On the contrary, when the content of a concept is articulated to a limited degree, the GoGAR is likely to be reduced to observational reports, pure descriptions, and naming of objects (Nilsson, 2020). For instance, simply responding by uttering "triangle" at the appearance of a triangle is pretty much what a parrot can be trained to perform, without the need of any inferential uptake. Hence, in expressing knowledge of a concept we are placing the concept in a language game, of justifying and being able to justify what we say (Sellars et al., 1997). In other words, we are placing it in a GoGAR in which we make claims, give reasons, ask for reasons and acknowledge, undertake and reject claims of others (Schindler & Seidouvy, 2019)

Although inferentialism is a theory that is attracting increasing attention in statistics education (Nilsson et al., 2017), it has the role of a background theory (Mason & Waywood, 1996). As a background theory, it provides a conceptualization of meaning and understanding (cf. Bikner-Ahsbahs & Prediger, 2010). Inferentialism, however, is not restricted to a specific subject-matter. The content that is coming into play or being expressed in a GoGAR is an empirical question. In the present study the focus is on statistical inference, and specifically on IHT. On this account, a conglomerate of relevant

[1] Brandom makes no significant distinction between *terms* and *concepts,* so in the rest of the paper *concepts* is used.

results from previous research in statistics education constitutes foreground theory (Mason & Waywood, 1996) in the present study.

2.2. RELEVANT RESEARCH ON STATISTICS EDUCATION

Sophisticated statistical reasoning is characterized by the integration of concepts and the ability to attend to and coordinate multiple aspects of the data being examined (Langrall et al., 2017). Taking into account this complexity, Makar and Rubin (2009) suggest organizing informal statistical inference around three key characteristics:

- A generalization or claim that extends beyond the data;
- Use of the data as evidence; and
- Expression with probabilistic (non-deterministic) language.

The first characteristic, generalization beyond data, contrasts statistical inference with descriptive statistics, which describes available data. Generalization beyond data entails an extension from the data at hand to a larger population, another sample, or the mechanism/process/context that generated the sample (Makar & Rubin, 2014). The second characteristic, use of data as evidence, recognizes that not all inferences are statistical. To be a statistical inference, it must be based on data rather than anecdote or deterministic causes (Makar & Rubin, 2014). Lastly, the third characteristic, probabilistic language, is a means of expressing uncertainty about the inferences drawn from data as evidence. Informally, students can use words such as “probably”, “maybe”, “unlikely”, “it could be”, to express, informally, degrees of uncertainty in a statistical inference (Makar et al., 2011).

In searching for evidence in data, Konold and Pollatsek (2002) argue for coming to see data as a mixture of signal and noise; that is, to see reasoning about data as the investigation of noisy processes that have a signature, a signal, that can be detected in sufficient data. The signal refers to patterns, to stable properties of a variable system—group features such as mode, mean and median—while noise refers to random variation (Konold & Pollatsek, 2002). Statistics education has traditionally focused on central tendency, but more recently variation has gained increased attention as a core idea of statistical reasoning (Pfannkuch & Wild, 2004; Shaughnessy, 2007). Ways of reasoning about variation include: (1) making comparisons within a data set, (2) making comparisons between data sets, and (3) making inferences from a given data set regarding an expected theoretical distribution (Mooney, 2002; Stohl Lee et al., 2010).

Several studies indicate students may find it difficult to understand the connection between outcomes and samples. For instance, when faced with data, learners may argue that a sample provides complete information about the population (seeing only a signal and no noise) (Ben-Zvi et al., 2012; Rubin et al., 1990), or that one should expect little or no variation in samples from an equiprobable distribution, even with a small sample size (Canada, 2006; Green, 1983; Watson & Kelly, 2004). On the contrary, if variation is stressed too much, students may believe that the sample provides no information about the population (seeing only noise and no signal) (Ben-Zvi et al., 2012). Shaughnessy and Ciancetta (2002) found that many students are willing to accept wide variation across several samples and maintain a belief that events are equally likely even with contrary visual and numerical evidence.

Hypothesis testing requires the drawing of inferences on whether variation can be understood as being caused by a chance effect, or if it is more likely that the variation is caused by a real difference in the population or in the underlying probability distribution. It concerns distinguishing signals from the underlying probability distribution in order to confirm or reject a hypothesis about the underlying probability distribution in question (Konold & Pollatsek, 2002; Zieffler et al., 2008). What Pratt et al. (2008) found, however, was that while students seemed to be searching for stable properties, invariance, across samples in a probability context, they were often puzzled as they were more responsive to the variations (noise) of the system. Studies also show that students can come to overemphasize contextual reasons when explaining or predicting data from an underlying probability distribution (Makar & Rubin, 2009). In Nilsson et al. (2018), for instance, students turned to material and *idiosyncratic explanations* (Watson et al., 2007) rather than to probabilistic and data-centered reasons when asked to predict or explain the distribution of outcomes obtained from a random generator, simulated by a bottle with colored marbles[2].

Stohl Lee et al. suggested that “if we want to promote attention to invariance across empirical distributions, then perhaps it can be productive to use contexts in which theoretical distributions are significantly different from a uniform distribution” (2010, p. 90). In studies conducted by Nilsson (2007, 2009), students were engaged in a dice game based on the total of two dice. The game consisted of a system of four setups of dice, and the studies showed how a systematic change between the setups encouraged the students to turn away from contextual or idiosyncratic reasoning in favor of more probabilistic and data-centered reasons when predicting and explaining frequency distributions. Pratt (2000) and Stohl Lee et al. (2010) showed how students can come to appreciate data-based evidence, and particularly the role of sample size, when engaged in computer simulations. In the setting of the Chance-Maker microworld, Pratt (2000) identified a *large number resource*, which involves an understanding that (relative) frequencies stabilize as a function of an increasing amount of data.

Drawing reliable informal statistical inferences requires controlling for sampling bias (Makar & Allmond, 2018). A statistical inference is never stronger than the data on which it is based, and it is critical that students develop an understanding of statistical inference that centers on the relationship between the legitimacy of the conclusions drawn from the data and the soundness of the process by which the data were generated (Cobb & McClain, 2004). Developing such an understanding, however, has proven to be difficult. Watson and Moritz (2000), for instance, showed that although students developed an understanding of variation and the role of sample size, they still failed to identify sample bias. Meletiou-Mavrotheris and Paparistodemou (2015) also noted that students’ responsiveness to contextual features interfered with their ability to account for bias in data production. In the present study, data production will be used as a generic term for processes of generating, collecting, and logging data.

3. THE STUDY

A small-scale teaching experiment over three lessons in a Swedish fifth-grade class (11–12 years old) was conducted in order to explore IHT in a probability context. The students were asked to act as data detectives (Shaughnessy, 2007), in the mission of distinguishing a non-uniform probability distribution from uniform probability distributions. The task was aligned with those suggested by Zieffler and colleagues (2008) for engaging students in informal inference by eliciting a judgment about which of two competing models or statements is more likely to be true.

3.1. CONTEXT, PARTICIPANTS, AND DATA COLLECTION

The students were familiar with traditional textbook teaching in mathematics, mixed with elements of group work and whole-class discussions. They were not accustomed to experimentation and had not been formally introduced to probability. Their teacher informed the researcher, the author of the paper, that the students had minor experience with decimal numbers, fractions, and percentages. The researcher designed all activities and tasks, together with a research colleague. In order to strengthen the study’s practical relevance and authenticity, the teacher was in charge of the teaching. The aim of the study, however, was not to examine the power of a certain teaching design but rather to explore students’ IHT. So, if the researcher saw an opportunity, he encouraged the teacher to challenge the students on certain issues.

Data were collected using four cameras. One camera, used for capturing whole-class discussions, was connected to a wireless microphone that was placed on the teacher. With this wireless microphone, the camera also recorded the teacher’s interaction with students during group work. The other three cameras were used for observing the group work. Each group camera had an external microphone placed on the group table.

[2] The bottles in Nilsson et al. (2018) are similar to those used in the teaching experiment in the present study.

3.2. RATIONALE FOR THE TEACHING EXPERIMENT

The teaching consisted of three lessons and was based on an activity in which a bottle containing a number of small, colored marbles was used as a random generator (Brousseau et al., 2001). Each lesson lasted about 60 minutes.

Lesson 1: Playing the color-run In Lesson 1 the class played four games of color-run. The game consisted of a bottle filled with marbles (Figure 1) and a game board (Figure 2). After shaking the bottle, the color of the marble that lies in the neck of the bottle is recorded. The color first recorded seven times wins the game. Three games were prepared in advance. A fourth game was initiated by the teacher and the students at the end of Lesson 1. In Lesson 1 the bottles were transparent. The contents of the bottles varied between the games, and the change was visible to the students on the game board. The aim of the design was to support reasoning about random variation and the relationship between underlying probability distribution (the sample space) and outcome frequencies. In the fourth game, the bottle was filled with one blue, four red, and four yellow marbles. The teacher turned the bottle around in all four games. One student was asked to log the observations on the game board, which was projected on the whiteboard at the front of the classroom. The teacher walked around in the classroom, asking the students to voice observations of an outcome.

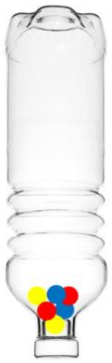
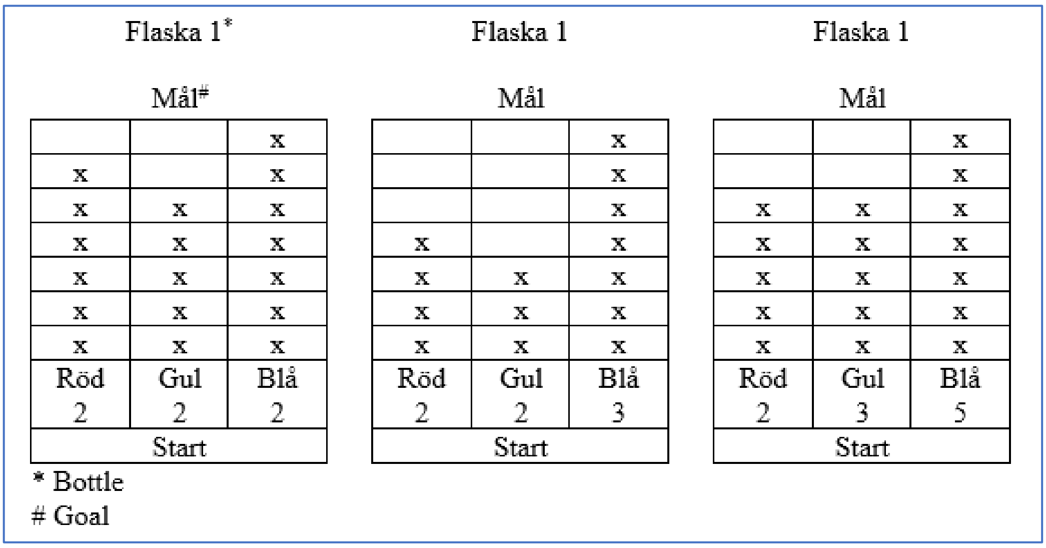


Figure 1. Bottle and marbles.



Flaska 1*

Mål#

		x
x		x
x	x	x
x	x	x
x	x	x
x	x	x
x	x	x
Röd 2	Gul 2	Blå 2
Start		

Flaska 1

Mål

		x
		x
		x
x		x
x	x	x
x	x	x
x	x	x
Röd 2	Gul 2	Blå 3
Start		

Flaska 1

Mål

		x
		x
x	x	x
x	x	x
x	x	x
x	x	x
x	x	x
Röd 2	Gul 3	Blå 5
Start		

* Bottle
Goal

Figure 2. A record of the results from the first three games. The bottles of the three games were prepared in advance, and their contents is shown under the colors: Röd = Red, Gul = Yellow, Blå = Blue

Lesson 2: IHT in the context of a data detective mission Lesson 2 began with the class discussing their game experience from Lesson 1 and formulating a theory on the relationship between the underlying probability distribution and the chance of a certain color winning a game. It was expected that this discussion would lay the foundation for the students' IHT in Lessons 2 and 3.

Next, the teacher introduced the data detective mission. She showed the students a bottle that was painted black. There was a little gap at the neck of the bottle so, when the bottle was turned upside down the students could observe only the color of the marble that ended up at this gap (Figure 3).

The data detective mission was described as follows: The color-run has been used in teacher education. When teacher-students play the game, it is assumed that all bottles contain an equal number of red, yellow, and blue marbles. There is a rumor, however, that in one bottle one of the colors is "on steroids"; that is, in one of the bottles there are more marbles of one color than the other two colors.

The students were told that their mission was to figure out which bottle contained a color on steroids (In short: *bottle on steroids*). The class was arranged in eight groups of three or four students each. Seven groups received a fair bottle, while one group unknowingly received the bottle on steroids.

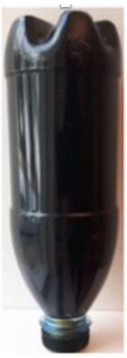


Figure 3. Covered bottle

Each group had their own worksheet with a table and diagram on which they were asked to log their observations (Appendix I). During the students' data production, the researcher and teacher noticed that the data might be biased for some groups, as they collected their data relatively quickly. So, after a short while, the teacher asked the students about their initial thoughts—whether they thought they had the bottle on steroids. The purpose of the break was also to attend to the process of the data production (Makar & Allmond, 2018). After the interruption, the teacher handed out new blank worksheets and the students were asked to start the data production from scratch. As the students had minor experience with fractions and relative frequencies, they were asked to log the absolute frequencies in their diagrams.

At the end of Lesson 2, the students were asked to decide whether they thought they had the bottle on steroids and to describe what motivated their decision. The students recorded their responses on a worksheet (Appendix II). In order to elicit probabilistic language and considerations (Makar & Rubin, 2009), they were also asked to determine how certain they were. In Lesson 2 the groups were not allowed to look at the data of the other groups: the focus of the task was within-sample investigations.

Lesson 3: Across-sample investigations Lesson 3 began with the teacher showing the diagrams from all groups (Figure 4), with the purpose of stimulating analyses across samples. The driving question in the across-sample investigations was to determine which of the samples were not the result of a uniform distribution: in other words, to determine which of the samples displayed a variation that was most likely caused by a genuine, real, difference in the probability distribution of the colors.

Four samples showed a difference that could have reasonably come from a bottle on steroids and not the result of a random process alone. Three of them were likely the effect of sampling bias. It was therefore decided that the teacher would encourage the class to carry out an extra round of samples, accompanied by a discussion about the importance of avoiding bias in producing samples (Meletiou-Mavrotheris & Paparistodemou, 2015). She emphasized, in particular, the importance of shaking the bottle properly between observations and ensuring that the students recording the observations did not fall behind the sampling process.

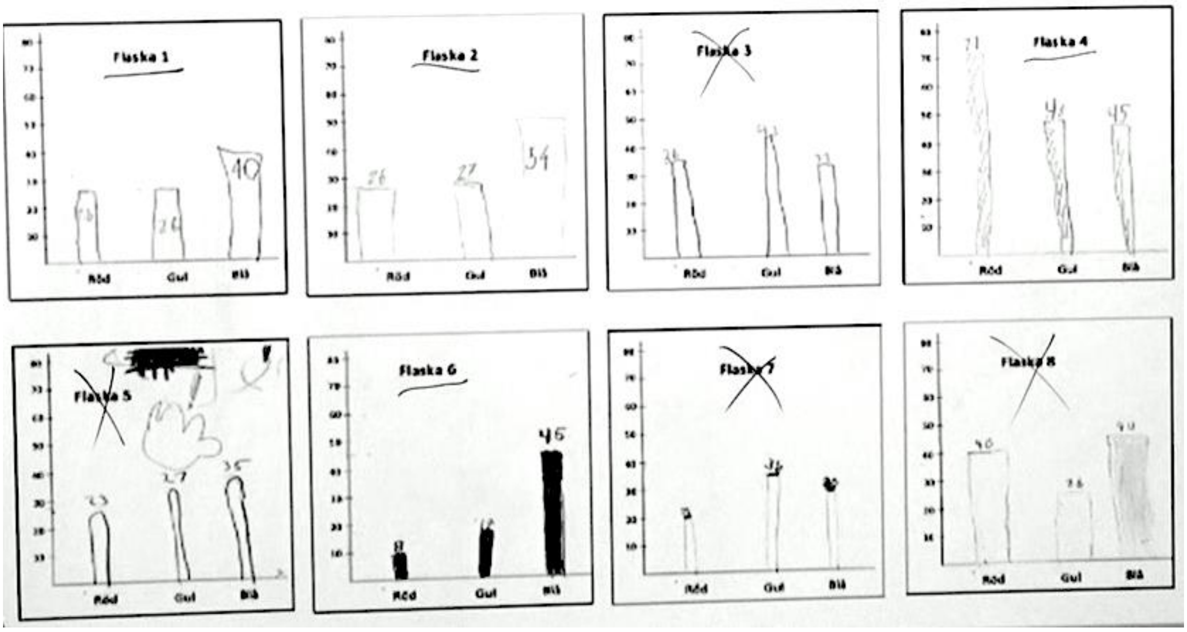
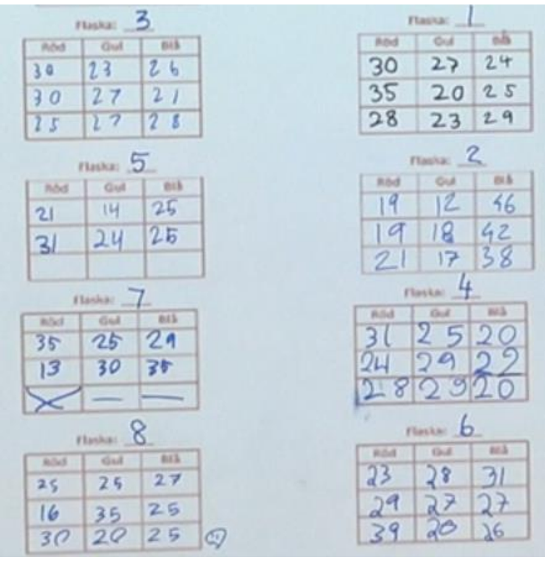


Figure 4. The results of all eight groups samples. Samples with high potential of being on steroids are underlined whereas bottles with low potential are crossed

In Lesson 2, one student suggested looking at whether there was a bottle in which one color won more often than the other colors in repeated color-runs. Because this idea was not part of the teaching the researcher had planned for in Lesson 2, the teacher did not follow up on this idea in the lesson. A modified version of the idea, however, was picked up in Lesson 3, with the students being asked to produce three new samples. The class agreed that each sample should comprise 80 observations.

Lesson 3 ended with a discussion based on the new samples (Figure 5), to decide which groups' samples were most likely not the result of a uniform probability distribution. The situation provided three overall patterns of sample analyses: comparing frequencies within each single sample of a group, across the three single samples of a group, and across group samples (cf. Mooney, 2002).


Flaska: 3

Röd	Gul	Blå
30	23	26
30	27	21
25	27	28

Flaska: 5

Röd	Gul	Blå
21	14	25
31	24	25

Flaska: 7

Röd	Gul	Blå
35	25	29
13	30	37
X	—	—

Flaska: 8

Röd	Gul	Blå
25	25	27
16	35	25
30	20	25

Flaska: 1

Röd	Gul	Blå
30	27	24
35	20	25
28	23	29

Flaska: 2

Röd	Gul	Blå
19	12	46
19	18	42
21	17	38

Flaska: 4

Röd	Gul	Blå
31	25	20
24	29	22
28	29	20

Flaska: 6

Röd	Gul	Blå
23	28	31
29	27	27
39	20	26

Figure 5. Results of the new samples in Lesson 3

3.3. METHOD OF ANALYSIS

This study is guided by the following research questions: How do students distinguish a non-uniform probability distribution from uniform probability distributions in a data-rich, experimentation-based learning environment, and what role do processes of data production play in their investigations? The analysis was based on the video recordings of the lessons and on students' written responses to the group tasks at the end of Lesson 2 (Appendix II). The analysis focused on the enacted GoGAR. This means that the analysis accounted for the GoGAR that was expressed in the classroom, regardless of which student was contributing. Through the enacted GoGAR, the analysis gives account of *how* students can express and make sense of IHT and what opportunities for learning are established within the frame of the GoGAR being enacted[3].

The analysis built on the assumption that humans are intentional creatures (Brandom, 1994; von Wright, 1971). In practice this means that, in making sense of students' ways of experiencing and dealing with the learning situation at hand, their behavior was interpreted as an act performed in order to realize a specific goal, a focal project, in the situation (Halldén, 1999; Nilsson & Ryve, 2010). Consequently, a focal project in a learning situation constitutes a major reason on which a GoGAR centers and is enacted. Makar and Rubin's (2009) framework of informal statistical inference was used to structure the analysis.

The analysis contained three steps. The first was to account for the focal projects in which the students became engaged. To count as a project relevant for IHT, it should concern a generalization or claim that extended beyond the data (Makar & Rubin, 2009) and be directed toward the underlying probability distribution. The question, "Do we have the bottle on steroids?" became a focal project. The second step was to examine the inferential relationships surrounding a focal project: what the project followed from and what followed from it, and what reasons were asked for and given (Brandom, 2000). A close look was taken at how the students came to use data as evidence (Makar & Rubin, 2009). On this account, the analysis was sensitive to how the students came to be responsive to contextual features in predicting or explaining random generated outcomes (Nilsson et al., 2018). The second step of the analysis was also guided by findings from research on how students distinguish signals in noisy processes (variation) (Konold & Pollatsek, 2002). The third and last step looked at relationships between the GoGAR episodes, to give an account of how IHT was enacted through the three lessons taken together.

4. RESULTS

The results are structured according to the GoGARs enacted along the teaching experiment. Each section includes a narrative with quotations taken from the transcripts of the video recordings and concludes with a commentary about the key ideas expressed. All participants' names are changed to pseudonyms. The teacher was given the name, Malin.

4.1. CONNECTING UNDERLYING PROBABILITY DISTRIBUTION TO OUTCOME FREQUENCIES

In Lesson 1, the blue color wins the three first color-runs. After the third run Malin asks, "Why did blue win?" Lena answers:

Lena: Because there was more blue in the bottle.

Some students, however, seem to think there is something special about the blue color, expressing idiosyncratic (Watson et al., 2007) reasons such as "God helped" and "Blue fought the most." Challenging such reasons, Malin initiates a fourth game, asking the students how many marbles of each color they should add to the bottle for this game. The class agrees to fill the bottle with one blue, four

[3] *Enacted* in enacted GoGAR can be compared to how Marton, Tsui, Chik, Ko, and Lo (2004) refer to *enacted* in relation to an enacted object of learning.

yellow, and four red marbles and then articulates the focal project: to test whether blue also wins this time.

The red color wins Game 4. Five yellow and no blue marbles are observed. Malin gives the students a few minutes to discuss in their groups why a certain color won. Again, some students voice idiosyncratic reasons, but do it with a smile or followed by a laugh. Alex, though, says:

Alex: The first time [Game 1] it was just luck that blue won … and then it was because they had more. And then, at the end [Game 4] red had the luck.
Malin: Why do you think blue had the luck in the first round? … Why could we say that blue had the luck in the first round, Elina [Elina is in the same group as Alex]?
Elina: It was two of the same kind.

In her response, Hallie specifies the inferential relationship between the underlying probability distribution in terms of the sample space, frequency distribution and random variation in terms of the winning color:

Hallie: It was the same probability for all to win, because there were equally many of each color.

Encouraged by Malin, Elina then continues to give reasons for why blue won the second and third games, in relation to the underlying probability distribution.

Elina: Because they had the highest probability, there were more marbles of the color blue.

Malin confirms Elina's explanation by revoicing, "More colors of the color blue than of the other colors." Their time is running out, and Malin decides to continue the discussion in Lesson 2, with the focal project of formulating an explicit theory on the relationship between the underlying probability distribution and outcome frequencies.

At the beginning of Lesson 2, Malin initiates the focal project of explicating the relationship between the underlying probability distribution and the chance of winning a color-run. She takes her departure in the GoGAR enacted in Lesson 1, regarding why blue won the first three games. The class negotiates on the theory:

> *If there are more marbles of one color, then the chance is greater that this color will win.*

The theory comes in an if-then structure, that is, it has the inferential form of a *conditional*. It is holistic in that it encompasses the basic ideas of underlying probability distribution, level of uncertainty, and mode as a measure of center (signal). The underlying probability distribution is expressed in terms of the contents of the bottles, and the mode refers to the color that wins. According to the level of uncertainty, it is also important to note that the conditional takes into consideration random variation. In other words, the inference does not simply move from, "If there are more marbles of one color" to "this color will win." Instead, it moves from, "If there are more marbles of one color" to "the *chance* is greater that this color will win." So, the theory is not formulated in deterministic but rather probabilistic terms, bringing probability language into an inferentially articulated generalization (Makar & Rubin, 2009).

4.2. CONTRASTING HYPOTHESIS TESTING WITH PARAMETER ESTIMATION

Malin introduces the data-detective mission in the second part of Lesson 2. The game experience has made explicit to the students that the more marbles of a color in the bottle, the greater the chance they will observe this color in a sample. The overall focal project in Lesson 2 is to infer whether a sample is the result of a bottle containing a uniform or a non-uniform probability distribution. Hence, the project is not about parameter-estimation, in terms of determining the proportions of colors in a bottle. That students without formal education in statistical inference can reflect on the difference between hypothesis testing and parameter estimation is shown in how they help to specifying the mission according to the principles of hypothesis testing. Alex asks:

Alex:	Should we calculate how many [marbles] there are of every color, or just [find] the one [bottle] that has the most [of one color]?
Malin:	We should see if we can find one bottle that constantly seems to give more of one color.
Alex:	So, we shouldn't calculate how many of each?
Malin:	No, we shouldn't calculate how many marbles there are inside [the bottles]. Good question!

Hallie asks for further specification and clarification:

Hallie:	Is it right that, all the other bottles that aren't on steroids, do they have equally many of all?
Malin:	All the bottles that don't have extra, they have equally many. There are eight groups; seven bottles have exactly the same contents, but one bottle has slightly more of one color.

These episodes show how students can be involved in a GoGAR by acts of contrasting. Alex and Malin make explicit the underlying idea of hypothesis testing by contrasting it with parameter estimation. Such contrasting speaks to what Brandom (1994) refers to as a negation inference: You increase your understanding of something when you come to understand what this something is not. The dialogue between Hallie and Malin then helps to specify the intended hypothesis testing further, by contrasting how one bottle is different from all the others.

4.3. COMPARING THE MODE WITH THE SECOND MOST OBTAINED OUTCOME

The groups begin to produce their samples. After a short time Malin interrupts the class, asking the students about their initial ideas about controlling for sampling bias (Meletiou-Mavrotheris & Paparistodemou, 2015). The groups then continue to collect their samples. At the end of Lesson 2 they draw an inference as to whether they think they have, or do not have, the bottle on steroids. They are also asked to state how certain they are about their inference (Appendix II). This section continues with a presentation of an overview of the IHT the groups formulated in their group work. The picture is then enriched by looking first at the discussion that arose at the interruption, and then at how the groups reasoned in their group tasks at the end of Lesson 2.

The mode, and its relation to the other outcomes, becomes a signal for all the groups. In all groups, the students focus on differences in frequencies, regardless of whether they infer that they have the bottle or not. In particular, they base their conclusions on the difference between the outcome obtained the most (i.e., the mode) and the outcome obtained the second-most. If the difference between the most and the second-most is large enough, they claim they have the bottle. In the overview of the students' IHT (Table 1), we see that many groups think they have the bottle on steroids. It is also noticeable that their level of certainty seems to be a function of the range of difference. More concretely, in Groups 2, 4, and 6 the difference between the mode and the second-most obtained outcome is 26 or more. All three of these groups are quite sure (over 80%) they can reject the possibility that their bottle contains a uniform distribution of marbles. Groups 1 and 5, who observe a difference of 11 and 8 respectively, are not very certain their sample is the result of a uniform distribution (null hypothesis).

Table 1. Overview of the groups' samples and inferences in Lesson 2

Group	Frequencies	Difference	Decision	Level of certainty	Reasons
2	26 (R), 27 (Y), 54 (B)	27	Yes	80%	There is a big difference between blue and the other colors.
4	71 (R), 40 (Y), 45 (B)	26	Yes	88%	There are almost twice as many red as yellow and blue.
6	8 (R), 17 (Y), 45 (B)	28	Yes	90%	We have a lot of blue compared to red and yellow.
1	26 (R), 26 (Y), 40 (B)	13	Yes (Maybe)	65%	We have a rather big difference.
5	23 (R), 27 (Y), 35 (B)	8	Yes (Maybe)	50%	We thought we had quite a lot of blue.
3	36 (R), 43 (Y), 33 (B)	7	Yes (Maybe)	62%	Their difference is small compared to the other groups.
7	21 (R), 36 (Y), 30 (B).	6	No	82%	No big difference between the numbers.
8	40 (R), 26 (Y), 44 (B)	4	No	57%	It's very even between red and blue.

The mode and differences in frequencies are central to the students' reasoning. This is in line with the theory the class agreed on in the introduction of Lesson 2. When Malin interrupts the data production, she implicitly gives further reason for this focus by stressing which color is ahead of the others:

Malin: I'm really curious how it looks. Take a quick look and see which color is leading in your group.

Each group tells her which color is leading in their sample. Different colors are leading in the different groups. Malin asks if any group already believes they know if they have the bottle on steroids. Three groups say they do. Malin reminds them that just one bottle is on steroids, and continues:

Malin: How then can we find out which [bottle] it is, if it's these, one of these three? What should we do to know? Joe?
Joe: You check how much there are leading to the one that's second, the one that got second-most.
Malin: You check the number we've got now. [...] We could count. Yes, what more could we do to be sure?
Aron: I know, we can ask everybody what color won, and the one who got the most points I think is the one.
Malin: Yes, we asked, and it seemed that both red and yellow – no, sorry, blue and red – won between the groups. [...] It was three, four of you to start with who thought you had the bottle on steroids. How would we make sure we know who has it?

Hallie responds by following up on Joe's idea, focusing on the difference between the most obtained (mode) and second-most obtained outcomes:

Hallie: You could check how big the difference is between the other one and the one that's leading. If there's only one point between them, then you may not be as sure as if it's, for instance, six points.

Malin makes the mode central to the GoGAR by turning the attention to what color is leading in the groups' samples. She says nothing, however, about the difference between the mode and the other outcomes; seeing a signal in the difference between the mode and the other outcomes is something that comes from the students. Also coming from the students is the idea to use this signal to give reasons for drawing an inference about which one of several proposed bottles likely does not contain a uniform distribution of marbles. Joe introduces the reason that it is not enough to just look at the mode. As a request for clarification, that is, an act of asking for further reasons, Hallie does not only repeat the general idea asserted by Joe but goes further to suggest they can look at the difference between the frequencies. Using concrete numbers, she extends and makes more explicit the enacted space of IHT, that the greater the difference between the frequencies of the mode and the other two outcomes, the greater the chance is that the sample is the result of a bottle on steroids. None of the other students react or add to Hallie's claims, and Malin does not continue to challenge the class by asking for reasoning regarding how they can solve the data-detective mission. In Table 1, however, we see the way Joe and Hallie reason is similar across all the groups.

Malin hands out a worksheet with a blank table and diagram (Appendix I), and the students begin producing new samples. We enter Lesson 2 again after all the groups have finished their samples[4] and Malin has handed out the questions that they are to discuss (Appendix II). Groups 7 and 8 use the observation of a small difference in frequency outcomes to give reasons for why a sample is likely the result of a uniform distribution. Group 7 comprises three students, of whom Hallie and Saga are two. This group received 21 (R), 36 (Y), and 30 (B). Saga claims that they have the bottle on steroids, while Hallie and the other student do not agree with this. Malin approaches Group 7 and asks Saga for her reasons:

Saga: Because there were a lot of yellow.
Malin: Aah [a tone that signals that she is not sure she agrees], if we look here then, 21, 36, 30.
Hallie: We were saying that it was very even.
Malin: You thought it was very even?
Hallie: When we did this [pointing to the theory on the board] it ended up like this also, and then there were equally many of each ... then it was even between red and blue, and yellow was a bit behind.

The group then agrees that they do not have the bottle on steroids, with 82% certainty.

Above we see how Hallie refers to a large difference in frequencies, to distinguish which bottle most likely contains a non-uniform distribution. Here, she extends the GoGAR on IHT by using the idea of difference in frequencies in the reverse case also; i.e., to infer why they cannot reject that a sample is obtained from a uniform probability distribution. Group 8 follows the same reasoning, using the observed small difference between the two most obtained outcomes as a reason for why they think they do not have the bottle on steroids. Group 8 received a sample with 40 (R), 26 (Y), and 44 (B). With 57%, certainty, they claim they do not have the bottle on steroids:

Mary: We think it's quite even between them [pointing to the piles of red and blue in the diagram].
Ellie: It's even between...
Mary: And that's why we can see that we don't have the bottle on steroids.
Ellie: Write down that it's even between red and blue.
Mila: Look, it's almost even between red and blue [pointing to the result in their table], so then we see that we don't have one color that has quite a lot because they're even.

Both Groups 7 and 8 show how students without formal education in hypothesis testing can express an understanding of random variation, which is central to hypothesis testing. This understanding includes knowing it is reasonable that there will be some difference between frequencies, even if the underlying distribution is uniform. With this understanding, the students do not find that the difference they have observed in their samples provides sufficient reason for rejecting the possibility that the observation is the result of a uniform distribution of colors. The behavior of random variation was further enacted in the GoGAR when the class compared between samples in Lesson 3.

[4] After the break, the groups restarted their data production from the beginning. It is the results of this second round of sampling that are presented in Table 1.

4.4. THE GREATER THE FREQUENCY DIFFERENCE, THE GREATER THE CHANCE OF A REAL DIFFERENCE IN UNDERLYING PROBABILITY DISTRIBUTION

Malin introduces Lesson 3 by projecting the diagrams of all the groups on the whiteboard (Figure 5):

> Malin: From this picture [Figure 4], these different diagrams, are we able to determine now, now that we know which bottle it is [which group has the bottle on steroids]? Do you think so?

The students come up with several suggestions. Leo, for instance, argues for Bottle 4 "Because, it [red] has 71 and it's leading by almost 30." And when Aron, a member of Group 3, sees the results of the other groups, he rescinds the suggestion that his group has the bottle on steroids, which they had claimed in Lesson 2 (Table 1). He justifies his changed opinion by claiming that their results are more even than those for Bottle 4. Aron expands on this frequency reasoning, also claiming that it is more likely that Bottle 2 is on steroids compared to Bottle 1:

> Aron: Red [pointing to the diagram of Bottle 1 on the whiteboard] and red on Bottle 2...they're equal. Yellow on both are almost equal. But then, blue on that one [Malin points to blue from Bottle 1] and blue on the other, then Bottle 2 has more.

In Table 1 it is noticeable that students seem to see the degree of certainty as a function of the size of the difference in frequencies. Aron's reasoning provides further empirical evidence of this. His focus is on the mode in the samples and, particularly, whether a mode is far ahead of the other colors. When the students explain which samples they consider to be least likely from the results of the bottle on steroids, they refer to the small differences between the outcomes in these samples. As expressed by Luke, "All [the frequencies] are around 30 and 20."

Two groups of bottles are distinguished: those with high potential of being on steroids and those with low potential. Malin underlines the samples the class thinks might come from the bottle on steroids and crosses out the ones they believe not likely to have come from that bottle (Figure 4). From a random perspective, this grouping is interesting in two respects. Firstly, it implies that the students understand how random variation works. They do not simply voice that the bottle producing the highest number of observations of one color is the one on steroids; by their grouping, they also consider that it will be frequency differences between samples, due to the random process. Secondly, however, the grouping implies that the students consider that randomness has only a limited power of explanation. Consequently, it is reasonable to say that the enacted space of IHT includes the central principle of hypothesis testing: that a small difference between samples can be the effect of randomness, while a large difference between samples most likely implies a real difference in the underlying process.

Based on the samples and their grouping, the students realize that it is difficult to claim with a high degree of certainty which bottle contains a non-uniform distribution of marble colors. Accordingly, it is not difficult for Malin to convince them of the need to produce more samples to get a stronger basis for their mission as data-detectives.

4.5. CONTROLLING FOR SAMPLING BIAS AND MODE DOMINANCE ACROSS SAMPLES

Before Lesson 3, the researcher makes Malin aware that the data collection in Lesson 2 was likely biased. For instance, the researcher knows that Bottle 4 contains a uniform distribution of marbles and, with such a distribution, the probability of obtaining one outcome 71 times or more out of 164 observations is less than 0.5%. So, before the students start collecting new samples, Malin discusses with them the importance of being careful when producing their samples. In the new data production, the groups are more careful and focused. Moreover, both Malin and the researcher are more active during the activity, reminding the students to be careful and focused. The results of the new samples are much more in line with the underlying probability distribution of each bottle.

The students are asked to produce three samples. The class agrees to collect about 80 observations in each sample. Before the students start, Malin asks them why it can be good to make three samples.

Alex answers, “The first time it might be just luck that one wins, but if the same one wins all three times then you’re almost certain.” Connecting to their focus on differences in outcome frequencies, Leo contributes to enriching the GoGAR by specifying that the three wins also need to be large, “If the same one wins really big every time, then you’re more certain.” By nodding and then revoicing student responses Malin confirms the essence of the idea that Alex’s and Leo’s reasoning implies comparing samples in relation to how dominant a certain outcome is (as the mode) across samples.

Figure 4 shows the whiteboard with all the new samples. Two groups perform only two samples. The focal project is to distinguish which group has the bottle on steroids; i.e. to distinguish which of the samples are not likely those of a uniform probability distribution. The class agrees on Bottle 2, reinforcing their focus on the mode and looking at the difference in frequencies in and across samples as reasons for their decision. In line with this, they also follow up on and compare the samples by looking at how dominant a certain outcome is as the mode across samples. That is, the fact that Bottle 2 was the only one with one outcome (blue) winning all three times provides the students further reason to infer that Bottle 2 contains a non-uniform probability distribution. However, the students do not only base their inference on the number of times an outcome appears as the mode; they also take into account how much the mode dominates compared to the other outcomes.

Various students claim that the results on the left side of Figure 4 are even and are thus from bottles with a uniform probability distribution of marbles. Malin then turns everyone’s attention to the results on the right side: the bottles the class was more unsure about, based on the samples generated in Lesson 2. Several students now immediately say that Bottle 2 is on steroids, and Hallie explains why:

Hallie: Because, yellow and red, they’ve been rather even...
Malin: 19, 12; 19, 18; 21, 17.
Hallie: …every time, and then blue has run ahead every time.

Looking a bit closer at what Hallie enacts, we see how inferentially rich her reasoning is. Firstly, it builds on the idea of difference in frequencies. It implies the insight that random variation can explain some frequency difference (the difference between yellow and red), but not when the difference becomes too large (the difference between the frequencies of blue and the other two colors). Secondly, it also includes the idea of mode dominance across samples, that the difference pattern from Bottle 2 is stable across all three samples. Extending the GoGAR in relation to how dominant an outcome is across samples, Malin talks through the samples resulting from Bottles 1, 4, and 6 to determine whether the same color from any of these bottles won every time. Unfortunately, this was not the case (Unfortunately, because such a situation could elicit further information as to how the students take into account the role of the size of mode dominance across samples). The students do come to express aspects of this, however, as Malin makes them reflect on the samples from Bottle 4:

Malin: Here, yellow won twice. What do we think of that then? Alex?
Alex: Maybe, eeh, that [the teacher holds her hand over the third sample] was a bit lucky, because all of them are rather even.

Mats adds:

Mats: But, from Number 2, blue won by far.

Before ending Lesson 3 the researcher asks the class why they think there was only a large difference in frequencies from one bottle this time. Linked to an increased understanding of the importance of controlling for bias (Watson & Moritz, 2000), several students offered the reason that they were more careful this time in shaking the bottle and in making a log of and counting their observations.

5. ANSWERING THE RESEARCH QUESTION

The aim of this paper has been to investigate IHT in a probability context through the research question: How do students distinguish a non-uniform probability distribution from uniform probability distributions in a data-rich, experimentation-based learning environment, and what role do processes of data production play in their investigations? The analysis outlines how young students with no formal education in hypothesis testing express a generalization beyond the data at hand, use data as evidence, and add probabilistic language to their reasoning (Makar & Rubin, 2009) in order to determine which of eight covered bottles does not involve a uniform distribution of red, yellow, and blue marbles. Taking the entire teaching experiment together, the analysis shows that the enacted IHT-related GoGARs involved:

- Connecting underlying probability distributions to outcome frequencies;
- Contrasting hypothesis testing with parameter estimation;
- Reflections on how reasonable it is that an observed difference in a sample is the result of a uniform probability distribution, and understanding that it is reasonable to obtain some difference even if the underlying probability distribution is uniform;
- Using the mode as a signal in data and, particularly, using the size of differences in frequencies as data-based reasons for rejecting a null hypothesis of an underlying probability distribution;
- The principle of hypothesis testing that a small difference across samples can be the effect of randomness, while a large difference across samples implies a real difference in the underlying process;
- Using probabilistic language indicating levels of uncertainty, and seeing the degree of certainty as a function of the size of differences in frequencies;
- The idea of mode dominance across samples; and
- A need to control for sampling bias.

6. DISCUSSION AND CONCLUDING REMARKS

This study provides new insights into informal statistical inference by showing how students can come to enact GoGARs on IHT in a probability context. The students who participated in the study had scant experience of fractions and, consequently, of the mean. In addition, they had never been taught the concept of relative frequencies. Despite this, the study shows how they were able to enact GoGARs on IHT: expressing an understanding of whether an obtained variation in data can be understood as being caused by a chance effect, or if it is more likely that the variation is caused by a real difference in the underlying probability distribution.

The study shows how students can distinguish the mode as a signal in data (Konold & Pollatsek, 2002), in order to infer on an underlying probability distribution. In the analysis it was evident that the mode is not an atomistic concept, used in some mechanical, procedural way by the students. It is given an inferential meaning in the GoGAR that was enacted throughout the three lessons. For instance, the mode was articulated as the color that leads, wins, or of which there is the most. The mode was also central in how the students came to focus on differences in frequencies and, particularly, in judging the underlying probability distribution based on the size of frequency differences. The mode also became central in GoGAR in the frame of mode dominance across samples. Taken together, the present study not only shows that the mode is intuitive for students to use in IHT, but also that focusing on the mode as a signal in data can have the potential to develop an inferential understanding of IHT.

This study also shows how across-data investigations from different random generators can challenge students to extend and develop their GoGAR on IHT. When the participating students had access to the samples of all the groups, they noted that there were several bottles that had the potential to contain a non-uniform distribution of color. Knowing that there was only one bottle with a non-uniform probability distribution, however, they realized the need for further investigation and, due to this, the need to sharpen the process of data production. From a teaching perspective it seemed important that only one color (outcome) differed, while the others were equally likely. This reduced the noise, making the signal of the mode clear and also making the signal of the difference between the mode and the other outcomes evident. This speaks to variation theory (Marton et al., 2004), noting that in order to discern some aspect of variation of a learning object, other aspects should be kept constant.

A well-known phenomenon in statistics education research is how contextual and idiosyncratic features can turn students' attention away from data-based reasoning (Makar & Rubin, 2009; Nilsson et al., 2018; Watson et al., 2007). Contrary to the students in Nilsson et al. (2018), the students in the present study did not fall as much into contextual reasoning for predicting or explaining obtained frequencies. The random generator used in the present study was similar to the one used in Nilsson et al. (2018) and the students were of the same age. How can the difference between the two studies be understood? The answer is likely to be found in how the games' design differed. In Nilsson et al. (2018) the students played with the same bottle in repeated games, whereas in the present study the underlying probability distribution was changed between each game and the contents of the bottles for each game were visible on the game board. This purposeful variation (Marton et al., 2004) between the games turned the students' attention to the contents of the bottles and thus supported them in privileging the underlying probability distribution in their GoGAR in order to make sense of the outcome frequencies of the color-runs.

The students privileging of data-based reasoning over contextual or idiosyncratic reasons does not mean that contextual features did not also play a central role in the present study. Indeed, they did! On this account, this study, specifically, reinforces the observed need for further research on the role of sampling bias (Cobb & McClain, 2004; Meletiou-Mavrotheris & Paparistodemou, 2015; Watson & Moritz, 2000). We begin to get a rather good picture of the difficulties students can have in understanding the role of sample size in order to determine probability outcomes. Pratt (2000) and Stohl Lee et al. (2010) also showed how students can come to realize the role of sample size when engaged in computer simulations. The present study does not underestimate research on how students reason about sample size, but this was not in focus here. The size of the samples was largely determined by the researcher and the teacher. Although the role of the sample size did not become a topic of the enacted GoGAR, the need to control for sample bias did. It became evident how important an understanding of sample production is for sophisticated and reliable statistical reasoning (Cobb & McClain, 2004). It does not matter how good students are at reading data if the data they are reading is biased. For instance, it would have had only a minor effect to let the students continue producing data in Lesson 2. Since the sampling was biased, the samples would most likely have been more biased! In other words, the fact that there is too much noise (Konold & Pollatsek, 2002) across the samples from Lesson 2 is not because the sample size was too small. The reason was that the sampling procedure was biased. On this account, the present study adds to the call by Cobb and McClain (2004), for further research on the role of generating, collecting, and recording data in data-based reasoning to justify whether or not to refute a null hypothesis (Stohl Lee et al., 2010).

ACKNOWLEDGEMENTS

This work was supported by the Swedish Research Council (dnr. 721-2012-4811). The author would like to thank Andreas Eckert for the valuable research assistance.

REFERENCES

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning, 13*(1–2), 5–26.
[Online: https://doi.org/10.1080/10986065.2011.538293]

Ben-Zvi, D., Aridor, K., Makar, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM - The International Journal on Mathematics Education, 44*(7), 913–925.
[Online: https://link.springer.com/article/10.1007/s11858-012-0420-3]

Bikner-Ahsbahs, A., & Prediger, S. (2010). Networking of theories: An approach for exploiting the diversity. In B. Sririman & L. English (Eds.), *Theories of mathematics education: Seeking new frontiers* (pp. 483–512). Springer.
[Online: https://link.springer.com/chapter/10.1007/978-3-642-00742-2_46]

Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.

Brandom, R. (1995). Knowledge and the social articulation of the space of reasons. *Philosophy and Phenomenological Research*, *55*(4), 895–908.

Brandom, R. (2000). *Articulating reasons: An introduction to inferentialism*. Harvard University Press.

Brandom, R. (2002). The centrality of Sellars's two-ply account of observation to the arguments of "Empiricism and the Philosophy of Mind". In R. B. Brandom (Ed.), *Tales of the mighty dead: Historical essays in the metaphysics of intentionality* (pp. 349-358). Harvard University Press.
[Online: https://philpapers.org/rec/BRATCO]

Bransen, J. (2002). Normativity as the key to objectivity: An exploration of Robert Brandom's articulating reasons. *Inquiry*, *45*(3), 373–391.
[Online: https://doi.org/10.1080/002017402760258204]

Brousseau, G., Brousseau, N., & Warfield, V. (2001). An experiment on the teaching of statistics and probability. *The Journal of Mathematical Behavior*, *20*(3), 363–411.
[Online: https://www.sciencedirect.com/science/article/abs/pii/S0732312302000780]

Canada, D. (2006). Elementary pre-service teachers' conceptions of variation in a probability context *Statistics Education Research Journal*, *5*(5), 36–63.
[Online: https://iase-web.org/documents/SERJ/SERJ5(1)_Canada.pdf?1402525006]

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 375–395). Kluwer.
[Online: https://link.springer.com/chapter/10.1007/1-4020-2278-6_16]

Derry, J. (2013). Can inferentialism contribute to social epistemology? *Journal of Philosophy of Education*, *47*(2), 222–235.
[Online: https://doi.org/10.1111/1467-9752.12032]

Doerr, H. M., Delmas, R., & Makar, K. (2017). A modeling approach to the development of students' informal inferential reasoning. *Statistics Education Research Journal*, *16*(2), 86–115.
[Online: https://iase-web.org/documents/SERJ/SERJ16(2)_Doerr.pdf]

Dolor, J., & Noll, J. (2015). Using guided reinvention to develop teachers' understanding of hypothesis testing concepts. *Statistics Education Research Journal*, *14*(1), 60–89.
[Online: https://iase-web.org/documents/SERJ/SERJ14(1)_Dolor.pdf]

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching*. Springer.
[Online: https://www.springer.com/gp/book/9781402083822]

Green, D. R. (1983). A survey of probability concepts in 3000 pupils aged 11–16 years. In D. R. Grey, P. Holmes, V. Barnett, & G. M. Constable (Eds.), *Proceedings of the first international conference on teaching statistics* (pp. 766–783). Teaching Statistics Trust.
[Online: https://iase-web.org/documents/papers/icots1/ICOTS1.pdf?1454729073]

Halldén, O. (1999). Conceptual change and contextualization. In W. Schnotz, S. Vosniadou, & M. Carretero (Eds.), *New perspectives on conceptual change* (pp. 53–65). Elsevier Science.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education*, *33*(4), 259–289.

Kula, F., & Koçer, R. G. (2020). Why is it difficult to understand statistical inference? Reflections on the opposing directions of construction and application of inference framework. *Teaching Mathematics and its Applications: An International Journal of the IMA*, *39*(4), 248–265.
[Online: https://doi.org/10.1093/teamat/hrz014]

Langrall, C., Makar, K., Nilsson, P., & Shaughnessy, J. M. (2017). Teaching and learning probability and statistics: An integrated perspective. In J. Cai (Ed.), *Compendium for research in mathemtics education* (pp. 490–525). National Council of Teachers of Mathematics.
[Online: https://www.nctm.org/Store/Products/Compendium-for-Research-in-Mathematics-Education/]

Makar, K., & Allmond, S. (2018). Statistical modelling and repeatable structures: Purpose, process and prediction. *ZDM - The International Journal on Mathematics Education*, *50*(7), 1139–1150.
[Online: https://link.springer.com/article/10.1007/s11858-018-0956-y]

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*(1-2), 152–173. [Online: https://doi.org/10.1080/10986065.2011.538301]

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105. [Online: https://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1).pdf#page=85]

Makar, K., & Rubin, A. (2014). Informal statistical inference revisited. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics* (pp. 1-6). International Statistical Institute. [Online: https://iase-web.org/Conference_Proceedings.php?p=ICOTS_9_2014]

Marton, F., Tsui, A. B., Chik, P. P., Ko, P. Y., & Lo, M. L. (2004). *Classroom discourse and the space of learning*. Lawrence Erlbaum.

Mason, J., & Waywood, A. (1996). The role of theory in mathematics education and research. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 1055–1089). Springer. [Online: https://link.springer.com/chapter/10.1007/978-94-009-1465-0_29]

Meletiou-Mavrotheris, M., & Paparistodemou, E. (2015). Developing students' reasoning about samples and sampling in the context of informal inferences. *Educational Studies in Mathematics, 88*, 385–404. [Online: https://link.springer.com/article/10.1007/s10649-014-9551-5]

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, *4*(1), 23–63. [Online: https://www.tandfonline.com/doi/abs/10.1207/S15327833MTL0401_2]

Moore, D. S. (2004). *The basic practice of statistics* (3rd ed.). W. H. Freeman.

Nilsson, P. (2007). Different ways in which students handle chance encounters in the explorative setting of a dice game. *Educational Studies in Mathematics*, *66*(3), 293–315. [Online: https://link.springer.com/article/10.1007/s10649-006-9062-0]

Nilsson, P. (2009). Conceptual variation and coordination in probability reasoning. *The Journal of Mathematical Behavior*, *28*(4), 247–261. [Online: https://www.sciencedirect.com/science/article/abs/pii/S0732312309000534]

Nilsson, P. (2020). A framework for investigating qualities of procedural and conceptual knowledge in mathematics: An inferentialist perspective. *Journal for Research in Mathematics Education, 51*(5), 574–599. [Online: https://www.jstor.org/stable/10.5951/jresematheduc-2020-0167]

Nilsson, P., Eckert, A., & Pratt, D. (2018). Challenges and opportunities in experimentation-based instruction in probability. In C. Batanero & E. Chernoff (Eds.), *Teaching and learning stochastics: Advances in probability education research* (pp. 51-71). Springer International Publishing. [Online: https://link.springer.com/chapter/10.1007/978-3-319-72871-1_4]

Nilsson, P., & Ryve, A. (2010). Focal event, contextualization, and effective communication in the mathematics classroom. *Educational Studies in Mathematics*, *74*, 241–258. [Online: https://link.springer.com/article/10.1007/s10649-010-9236-7]

Nilsson, P., Schindler, M., & Bakker, A. (2017). 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. [Online: https://link.springer.com/chapter/10.1007/978-3-319-66195-7_11]

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, thinking and reasoning* (pp. 17–46). Kluwer Academic Press. [Online: https://link.springer.com/chapter/10.1007/1-4020-2278-6_2]

Pratt, D. (2000). Making sense of the total of two dice. *Journal for Research in Mathematics Education, 31*(5), 602–625. [Online: https://www.jstor.org/stable/749889?seq=1#metadata_info_tab_contents]

Pratt, D., Johnston-Wilder, P., Ainley, J., & Mason, J. (2008). Local and global thinking in statistical inference. *Statistics Education Research Journal*, *7*(2), 107–129. [Online: https://iase-web.org/documents/SERJ/SERJ7(2)_Pratt.pdf]

Schindler, M., & Seidouvy, A. (2019). Informal inferential reasoning and the social: Understanding students' informal inferences through an inferentialist perspective. In G. Burril & D. Ben-Zvi (Eds.), *Topics and trends in current statistics education research: International perspectives* (pp. 153–171). Springer.
[Online: https://link.springer.com/chapter/10.1007/978-3-030-03472-6_7]

Sellars, W., Rorty, R., & Brandom, R. (1997). *Empiricism and the philosophy of mind*. Harvard University Press.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *The second handbook of research on mathematics* (pp. 957–1010). National Council of Teachers of Mathematics.

Shaughnessy, J. M., & Ciancetta, M. (2002). Students' understanding of variability in a probability environment. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society*. International Statistical Institute.
[Online: http://iase-web.org/documents/papers/icots6/6a6_shau.pdf]

Stohl Lee, H., Angotti, R. L., & Tarr, J. E. (2010). Making comparisons between observed data and expected outcomes: Students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal, 9*(1), 68–96.
[Online: https://iase-web.org/documents/SERJ/SERJ9(1)_Lee.pdf]

von Wright, G. H. (1971). *Explanation and underestanding*. Routledge and Kegan Paul.

Watson, J., Callingham, R., & Kelly, B. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning, 9*(2), 83–130.
[Online: https://doi.org/10.1080/10986060709336812]

Watson, J., & Kelly, B. (2004). Expectation versus variation: Students' decision making in a chance environment. *Canadian Journal of Science, Mathematics and Technology Education, 4*(3), 371–396.
[Online: https://doi.org/10.1080/10986060709336812]

Watson, J., & Moritz, J. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31*(1), 44–70.
[Online: https://doi.org/10.2307/749819]

Wittgenstein, L. (1968). *Philosophical investigations* (Vol. 3). Blackwell.

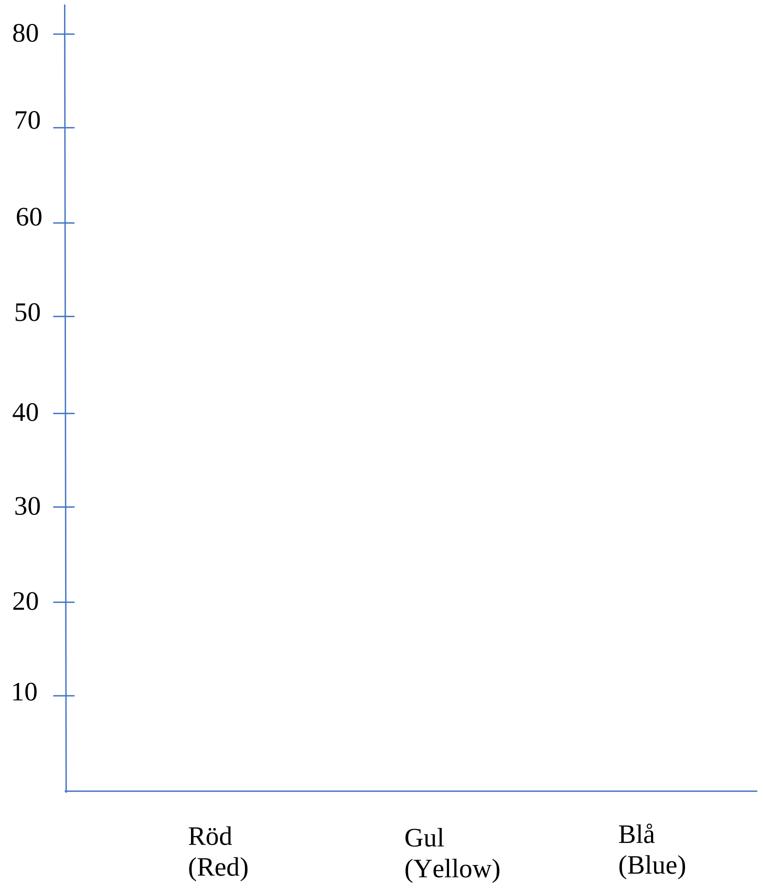
Zieffler, A., Garfield, J., Delmas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40–58.
[Online: https://www.stat.auckland.ac.nz/~iase/serj/SERJ7(2)_Zieffler.pdf]

PER NILSSON
Örebro University
School of Science and Technology
SE-701 82 Örebro
Sweden

Appendix I

The translation into English was not part of the original version.

Röd (Red)	
Gul (Yellow)	
Blå (Blue)	



Appendix II

1. Tror ni att det är er flaska som har den dopade hästen? (kryssa ett alternativ) (Do you think your bottle is the one with the color that is on steroids? (mark one option))

□ □ □

Ja (Yes) Nej (No) Kanske (Maybe)

Förklara hur ni tänker (Explain how you reason):

....

....

....

2. Hur säkra är ni på att ert svar stämmer? Markera med ett kryss på linjen:
(How certain are you that your answer is correct? Mark on the line with an X):

Inte alls säkra
(Not certain at all)

Helt säkra
(Completely certain)

Förklara hur ni tänker (Explain how you reason):

....

....

....

3. Om ni inte är helt säkra, finns det något ni skulle kunna göra för att bli mer säkra? (Ni får inte öppna flaskan.) Förklara:
(If you're not completely certain, is there something you could do to be more certain? (You're not allowed to open the bottle.) Explain:

....

....

....