HYPOTHETICAL LEARNING TRAJECTORY ON INFORMAL HYPOTHESIS TESTING IN A PROBABILITY CONTEXT

PER NILSSON
Örebro University
per.g.nilsson@lnu.se

ABSTRACT

A design experiment where students in Grade 5 (11–12 years old) play the Color Run game constitutes the context for investigating how students can be introduced to informal hypothesis testing. The result outlines a three-step hypothetical learning trajectory on informal hypothesis testing. In the first step, students came to favor sample space reasoning over idiosyncratic reasoning when the sample space was changed between color runs. In the second and third steps, students used degrees of variation in the distribution of the mode across samples to infer whether an unknown sample space was uniform. Students' reasoning disclosed the logic: the larger the variation, the greater the reason for rejecting a uniform sample space.

Keywords: *Statistics education research; Informal statistical inference; Hypothesis testing; Inferentialism; Reasons, Sample-space*

1. INTRODUCTION

This study draws on the increased attention on how learning environments can be designed to support young students' statistical reasoning (Ben-Zvi et al., 2018; Langrall et al., 2017). More specifically, it intends to contribute to the recognized need for further research to explore and characterize learning environments, which expose young students to situations that invite them to use informal methods in making statistical inferences (Makar & Rubin, 2009; Meletiou-Mavrotheris & Paparistodemou, 2015; Zieffler et al., 2008).

Inferential statistics is an important topic in statistics courses for many disciplines (Sotos et al., 2007). It is also a topic that has proved difficult to learn (Dolor & Noll, 2015; Krishnan & Idris, 2015) and many students enter introductory courses in statistics with misconceptions (Sotos et al., 2007). It is suggested that the learning of inferential statistics should start in the early years of schooling (Meletiou-Mavrotheris & Paparistodemou, 2015), with instructional support that allows students to engage in inferential statistics informally. In other words, there is a need to understand more of how to elicit and develop students' informal methods on inferential statistics to facilitate the learning of the formal methods (Makar & Rubin, 2009; Nilsson, 2020b).

Hypothesis testing is a main method of inferential statistics. It is a complex method, and it has proved difficult for students or teachers to fully understand the formal theory of hypothesis testing (Dolor & Noll, 2015; Zieffler et al., 2008). Informal statistical inference has received increased attention in recent years and studies show both younger students' difficulties and ability to make claims of a wider universe beyond data with informal methods (Langrall et al., 2017). However, studies usually take a relatively general perspective on informal statistical inference. There is less of detailed knowledge about the teaching and learning of informal statistical inference specific to hypothesis testing, that is, about *informal hypothesis testing* (Lee et al., 2010; Nilsson, 2020b).

Another observation is that informal statistical inference is most often investigated in computer-based learning environments, where students are offered the opportunity to easy produce and repeat samples and organize and analyze data with different forms of representations. However, there is a risk that the power of computers encourages technological "short-cuts", that may create undesired habits or gaps in understanding (Langrall et al., 2017). There is a need for balanced use of the affordances in technologies (Rubin, 2007). On this account, the aim of the present study is to contribute to educational

Statistics Education Research Journal, 22(2), Article 11. https://doi.org/10.52041/serj.v22i2.425

research on informal hypothesis testing by investigating a hypothetical learning trajectory (Simon, 1995) that takes departure from students' practical experiences of chance and random variation.

A design experiment (Cobb et al., 2003) in which students in Grade 5 (11–12 years old) play the *Color Run* activity constitutes the context of investigation. I provide details of the design experiment after presenting relevant research, defining informal hypothesis testing, and addressing the research question of the study.

2. THEORETICAL BACKGROUND

2.1. HYPOTHESIS TESTING

In hypothesis testing we make a guess about the value of a 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). Zieffler et al. (2008) express the development of hypothesis testing reasoning as the ability to demonstrate an "understanding of a *p*-value as an indicator of how likely or surprising a sample result, or a result more extreme, is under a certain hypothesis, and the action of rejecting this hypothesis if the *p*-value is small enough" (p. 45). The guess about the parameter is formally the *null hypothesis* and is most often the assumption that there is no difference, for example, between outcomes in a population or between the effect of different treatments. Hence, the method of hypothesis testing centers around the relationship between the expected and the observed distribution of samples (Lee et al., 2010).

2.2. REASONING ABOUT VARIATION

Drawing an inference on the relationship between the expected and the observed distribution of data requires an understanding of the role of the space of possible outcomes (the sample space) and random variation in the distribution of data (Dolor & Noll, 2015; Pratt et al., 2008) since, "if one has no prior expectation about what a result should be, then no result is unusual" (Thompson et al., 2007, p. 220). Students may recognize that variation should be expected but have difficulty knowing how much variability to expect (Biehler et al., 2018; Mooney et al., 2014). When faced with data, learners may argue that a sample provides complete information about the population (Ben-Zvi et al., 2012; Rubin et al., 1990), or that one should expect little or no variation in samples from a uniform 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 (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. However, if students have an incorrect model of the underlying probability distribution, it does not matter how good an understanding they have of random variation. For instance, in Nilsson (2007) students were exposed to non-uniform sample spaces, in the setting of a game, based on the total of two dice. When predicting the empirical distribution of possible totals, students initially confirmed the equiprobability bias (Lecoutre, 1992). It appeared, however, that the application of the equiprobability expectation was not based on any reflections that the number of ways in which each sum could be represented was different. When students noticed that different totals could appear in a different number of ways during the game, they turned to a non-uniform expectation of how data should distribute over the totals of two dice.

2.3. REASONING ABOUT THE EXPECTED AND OBSERVED DISTRIBUTION OF DATA

Students may find it difficult to consider sample variation in the connection between an expected distribution of samples and the underlying theoretical distribution (Van Dijke-Droogers et al., 2020). Ireland and Watson (2009) emphasize the use of relative frequency displays of data in helping students to make proper connections between expected and observed distributions of data. Using displays of relative frequencies, however, seems not to be easy or natural to many students. Lee et al. (2010), for instance, examined middle school students' informal hypothesis testing in the setting of a computer-simulated die-tossing experiment. The students coordinated their mental image of an expected distribution of data, based on an imagined (hypothesized) theoretical distribution, with an observed distribution of data. Only a few students used the relative frequency table provided, to confirm or refute an imagined theoretical distribution. Instead, and like students in Nilsson (2020b), students attended to and compared absolute frequencies within samples. For instance, that one outcome appeared only a few times, compared to other outcomes in the sample, was used as a reason to refute that a computer-simulated die was fair. In Nilsson (2020b) focus was on the mode; if an outcome was observed significantly more times than other outcomes in a sample, students inferred that the underlying, theoretical distribution was not uniform.

2.4. LEARNING WITH CONCRETE RANDOM GENERATORS

The students are supposed to generate data with concrete random generators in the present study. Working practically with random experimentation provides students the opportunity to sense random behavior (Fischbein, 1975). However, when using concrete random generators, one needs to be alert to students' tendency to appreciate contextual, material, and idiosyncratic explanations over statistical and data-centered reasons (Makar & Rubin, 2009; Meletiou-Mavrotheris & Paparistodemou, 2015; Watson et al., 2007). In Nilsson et al. (2018), students worked with a plastic bottle, like the one used in the present study (Figure, 1). Some students argued it matters how the bottle was shaken and turned around and other students argued that the order of how the different colors were put into the bottle would play a role. In Nilsson (2020b), however, students did not fall as much into contextual reasoning as in Nilsson et al. (2018), despite a similar bottle being used and students being at the same age. How can we understand this difference? In Nilsson et al. (2018) the students played with the same bottle in repeated games, whereas in Nilsson (2020b) the underlying sample space was changed between each game and visible on the game-board. The variation in sample space turned students' attention to the contents of the bottles and supported them in privileging statistical reasoning. The present study follows the approach of Nilsson (2020b).

2.5. DEFINING INFORMAL HYPOTHESIS TESTING

I adapt to the pragmatism theory of inferentialism (Brandom, 2000) as a background theory (Mason & Waywood, 1996) for conceptualizing teaching and learning in informal hypothesis testing. Inferentialism is a semantic theory of how people ascribe meaning to and develop an understanding of concepts. Inferentialism should not be confused with statistical inference, despite that inference is at the core in both cases. In statistical inference, inference refers to the act of making a claim beyond data. In inferentialism, inference refers to the act of distinguishing what follows from the applicability of a concept (a statement) and what it follows from.

In inferentialism, rationality is defined in terms of reasons. Understanding involves having a grasp of reasons, driven by the peculiar force of striving for more acceptable and better reasons (Bakhurst, 2011). For instance, when we make a claim beyond data, grasping the claim entails getting to grip with the conditions of its application. It is to know, in the practical sense of being able to distinguish, what follows from the applicability of the claim and what it follows from (Brandom, 2000). For the learning of statistics, this implies teaching where students are invited to make claims (take a stand) and are challenged to ask and give reasons for their claims, in explorative practices of experimenting with data, making connections, explaining, inferring, and generalizing (Bakker & Derry, 2011).

In the formal method of hypothesis testing, we focus on a parameter value, such as the mean, of one outcome, out of two or more possible outcomes. On an ordinary die there are six outcomes, but in the formal method of hypothesis testing we are focusing on only one of these. For instance, if we want to use the method to decide whether a die is biased, we can look at how many sixes we obtain in a sample of 60 throws. The other five outcomes are not explicitly available or considered. We ask then, how likely the result we obtain would be, if the die was fair, that is, under the assumption of a parameter value of 1/6. We apply a formal method for calculating a *p*-value, which is, the probability of obtaining a certain result or a result more extreme, under the null hypothesis.

In the present study, informal hypothesis testing is defined according to Lee et al. (2010) and Nilsson (2020b). This means that students are not supposed to focus on the simulated distribution of one outcome and of simulating or calculating a *p*-value according to this distribution. Instead, the focus is on how students use variations across samples to informally infer on how likely it is that a random generator is uniform, and for the action of rejecting a random generator as uniform. For instance, in Lee et al. (2010) the students did not look at how a single outcome of a die was distributed across repeated samples. Instead, they were asked about whether the outcomes of a die were equiprobable or not. That is, the students were exposed to the distribution over all six outcomes and reasoned about whether a die was fair according to how often an outcome was observed, compared to the other outcomes. This is in line with Nilsson (2020b), who observed students being engaged in informal hypothesis testing, by comparing samples, drawn from uniform and non-uniform random generators. The study showed that students found a large variation between outcomes in a sample result as a reason to infer that the sample was drawn from a random generator with a non-uniform sample space. Another way students compared outcomes in Nilsson (2020b) was to look at the distribution of the mode (modal value) across samples. Students expected that, if the outcomes are equiprobable, each outcome should be obtained as the mode about as often in a series of samples.

2.6. RESEARCH QUESTION

Given the need for research that explores and characterizes learning environments, which invite young students to use informal methods in making statistical inferences, the present study addresses the research question:

> How can students be introduced to use variations across samples to informally infer on how likely it is that a concrete random generator is uniform and for the action of rejecting a random generator as uniform?

3. METHOD

A design experiment (Cobb et al., 2003), where students in Grade 5 (11–12 years old) are playing the Color Run constituted the context for investigating a hypothetical learning trajectory (Simon, 1995; Simon & Tzur, 2004) on informal hypothesis testing, as an answer to the research question. A hypothetical learning trajectory consists of a learning goal for students, underlying learning processes, and the means (tasks, activities) used to support these processes and to achieve the learning goals.

3.1. THE RATIONALE OF THE DESIGN EXPERIMENT

The conjectured end-goal of the design experiment was that students should demonstrate an ability to use variations across samples as reasons to informally infer on how likely it is that a concrete random generator is uniform, and for the action of rejecting a random generator as uniform. The tasks included were aligned with those suggested by Zieffler et al. (2008) for engaging students early in statistical inference by eliciting a judgment about which of two competing models or statements is more likely to be true.

The tasks were designed by the investigator. Sara[1], the class teacher, ran the lessons and was supposed to act as usual. The students were familiar with traditional textbook teaching in mathematics, mixed with elements of group work and whole-class discussions. The students had a minor experience of descriptive statistics and no experience of inferential statistics. The student had been taught on rational numbers and fractions.

The hypothetical learning trajectory was structured over an instructional sequence of three lessons, each based on the Color Run (Nilsson, 2020b). Each lesson lasted for 60 minutes. It was one day between each lesson.

Lesson 1 – Promoting sample-space reasoning by changing sample space in Color Run. Color Run is comprised of a bottle filled with marbles, and a gameboard. 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 to Lesson 1 (Figure 1).

[1] Sara is a pseudonym for the teacher's real name.

Each game was structured in three phases: predicting-playing-reflecting. First, the students were asked to predict their expectations of which color they thought would win and to provide reasons for their predictions. Second, the game was played. Third, the students were asked to reflect on the outcomes of the game, connecting back to their predictions.

A principal idea of the variation theory is that we can only discern an aspect if we experience variation in that aspect (Runesson, 2006). Since the sample space was an aspect for students to discern, the sample space was changed among the three games. Color Run should also incite reflections on random variation. A long game would help students to discern the regularity of the sample space. The shorter the game, the greater the effect of chance. After playing the game of different durations when designing the game, it was decided to use a seven-step game board (Figure, 1). A list of reasons is to the left of Game 1. At the top of the list is "weight", then "smooth" and at the bottom "luck, chance". The row at the bottom of the three game-boards tells the colors—red to the left, yellow in the middle and blue to the right—and the number of each color in the bottle. At the very bottom, the number of votes from the students for each option is recorded.

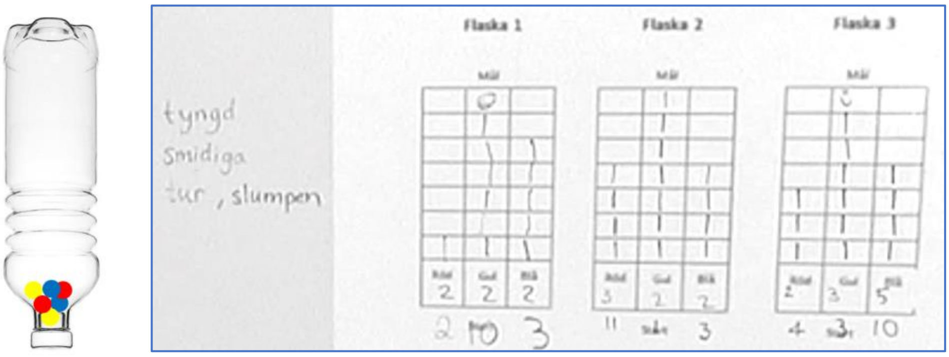

Figure 1. The transparent bottle and the three games, ordered from left to right.

Lesson 2 – Comparing samples from the same random generator. Lesson 2 was designed to support students to use variation between outcome frequencies across samples to infer on how likely it is that a random generator is uniform and for the action of rejecting a random generator as uniform. Eight bottles were prepared for Lesson 2: Seven with a uniform sample space of two red, yellow, and blue marbles and one with a non-uniform sample space of four blue, two red, and two yellow marbles. The shape of the bottles was the same as in Lesson 1. However, this time the bottles were covered. They were sprayed in black, with only a small transparent slip at the neck of the bottle. When the bottle was turned upside-down, it was possible to see only the color of one marble.

Students were grouped in pairs. The students were told that seven bottles were fair and that it was more marbles of *one* of three colors in one bottle. The class referred to the unfair bottle as the bottle on steroids. To avoid reasoning on parameter estimation, the students were not told how many colors there were of each color in the bottles. Following the logic of hypothesis testing, the target of the learning trajectory was to decide on which of the bottles are on steroids, i.e., which of the bottles does not have a uniform sample space. The groups played Color Run eight times and were then asked to draw an inference of if they thought they had the bottle on steroids or not based on their eight games. The groups were not allowed to look at the other groups during the group work.

In Lesson 1, each game was seven steps. The yellow color won all three times, even though there were fewer yellow marbles compared to red marbles in Game 2 and blue marbles in Game 3. To increase the chance that the bottle on steroids would stand out, but still maintain uncertainty, it was decided to extend the game board to ten steps. Since the games are still rather short, and there are only eight games, there is a relatively good chance for large variations. In Lesson 2, several groups also came to claim they had the bottle on steroids (see Table 1 below). Sara used the situation that several bottles were suggested as candidates for the bottle on steroids, to motivate further investigations in Lesson 3.

Lesson 3 – Comparing samples from different random generators. In Lesson 3, students were supposed to compare outcome frequencies from the different bottles, to find out the bottle with a non-uniform sample space, that is, with an unequal number of blue, red, and yellow marbles. The lesson was structured in four phases. First Sara handed out the game boards from Lesson 2 and asked the students to reflect on if they keep to their decision and explanation. Next, Sara gathered all groups in a whole-class discussion, where the class negotiated on candidates for bottles on steroids, by comparing the samplings from all eight bottles. In the third phase, the students were arranged into four larger groups, each given one of the four candidates. They received a copy of the eight games from Lesson 2 and a blank game board to play eight new games with the bottle they were assigned. In the last fourth phase, the class compared the samplings, each of 16 samples (games), from the four bottles. They inferred on which bottle was on steroids. Sara also asked the students if they thought any other bottle was on steroids. With this question, she challenged the students further on what they consider enough variation for rejecting, or not rejecting, that a random generator is fair, that is, has a uniform sample space.

3.2. DATA COLLECTION AND ANALYSIS

Data were collected using four cameras. One camera, capturing whole-class discussions, which was connected to a wireless microphone 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 group work. The groups' worksheets from Lessons 2 and 3 were also collected.

The analysis followed the principle of an inferentialist analysis (Brandom, 2000; Nilsson, 2020a), in relation to the analysis of a hypothetical learning trajectory (Simon, 1995). This means that the analysis was focused on students' reasoning and the instructional means that supported their reasoning. Inferentialism helps us understand that it is reasonable to assume that behavior and speech acts are rational in terms of reasons (Brandom, 2000). Hence, to understand and make sense of a claim, means to understand what the claim follows from, the reasons for the claim, and what follows from the claim, and how the claim is used as reasons for other claims. In the present case, I searched particularly for how students used and were supported to use, variation across samples to formulate informal reasons for inferring on how likely it was that a random generator was uniform, and for the action of rejecting a random generator as uniform.

4. RESULTS

The analysis resulted in a three-step hypothetical learning trajectory on informal hypothesis testing. Each step relates to the three subsequent lessons of the design experiment.

4.1. MOVING FROM MATERIAL REASONS TO SAMPLE SPACE REASONS

To understand what to expect in the distribution of data, students needed to have an insight of the relationship between the underlying sample space and data distribution. Supporting students on this insight was the objective of the three games of Color Run in Lesson 1. However, what we see in Figure 1 is that yellow won all games, despite the favor of red in Game 2 and blue in Game 3. Did this result ruin the objective of the three games? It did not! The game activity supported students to make the connection that, even if the number of favorable outcomes is beneficial for a color to win, it does not determine a game; the game is also a matter of chance.

Material reasons were most evident in the prediction phase of Game 1 and in explaining the results of Game 2 and 3. When asking the students for their votes in Game 1, Sara held the bottle horizontally in front of the class. The students took notice of that a yellow marble was at the front of all marbles and used this as a reason for why yellow would have an advantage. Sara did not shake the bottle to challenge this idea. The result of Game 2 and 3 did not match the proportions of the sample spaces. Most students accepted this as being a matter of chance, but some gave material reasons like, "Yellow was at the front," and "Maybe they are heavier, they look bigger." Sara posted the suggestions on the whiteboard (Figure 1). When reflecting on the result of Game 1, Sara challenged the idea that yellow won, because it was at the front. She did this by shaking the bottle and showing it is impossible to tell which marble is at the front. After this practical confrontation, the class agreed on removing a color's position in the bottle as a reason.

How then did the game support students' ability to connect sample space and data, and to take into consideration random variation in this connection? In predicting the expected result of Game 1, connections between sample space, data, and random variation were articulated in terms of the idea that it does not matter which color to vote for. Mathew, voting on blue, explained, "I don't know. One should pick one and it doesn't matter which one." After Game 1, several students also referred to luck as a reason when explaining why yellow won.

In predicting Game 2 and 3 most students showed evidence of sample-space reasoning. Daniel, voting on red in Game 2, claimed, "It is most of them." Similarly, Claire voiced, "I voted for blue, since they are most," when predicting Game 3. From the video we see that Sara took time to make the students aware of how the sample space differed from the previous game(s). The variation in sample space between the bottles was also visible on the whiteboard at the front of the class (Figure, 1). These actions and features of the teaching were critical in supporting students to engage in sample-space reasoning.

The relationship between Game 2 and 3 supported students to connect sample space and chance. For instance, several students explained the win of yellow in Game 2 as a matter of chance. And, even if Game 2 could support that it does not matter what color you vote for (everything is just a matter of chance) this did not happen. Instead, when predicting Game 3, most students again turned to the most favorable outcome to give a reason for why they voted for blue. When yellow again won, the same students referred to chance to give reasons for this result.

4.2. EXPECTING SMALL DEVIATIONS IN THE MODE FROM A UNIFORM SAMPLE SPACE

Eight bottles were prepared for Lesson 2: Seven with a uniform sample space of two red, yellow, and blue marbles and one with a non-uniform sample space of four blue, two red, and two yellow marbles. In Lesson 2 the students should act as data detectives (Shaughnessy, 2007), to figure out which bottle was on steroids, i.e., which bottle had more marbles of one color.

Sara began by asking the students to reflect on why most of them voted on red in Game 2 and on blue in Game 3 in the previous lesson. She did this to push on students' ability to connect sample space, variation, and data distribution. Mathew showed evidence of the gambler's fallacy, arguing, "I voted on blue in the last [game] also because I thought that, now that yellow had won two times, it might be not that probable it wins also the third time." Even if other students had the same opinion as Mathew, most students articulated they voted on red in Game 2 and blue in Game 3, because it was more red marbles in Game 2 and more blue marbles in Game 3. Sara was neutral to all suggestions. Applying a non-evaluative attitude, she encouraged the student to shape their own understanding and to test and develop their understanding in the subsequent inquiry.

Sara introduced the data-detective mission and the class talked about that a data-detective search for evidence in data. Each group was given a worksheet with eight games, each ten steps long. After playing the games, they were asked to decide if they thought they had the bottle on steroids or not and how certain they were (Appendix, 1). Three groups inferred they had the bottle on steroids and five inferred they did not (Table 1). Besides Group 4 and Group 5, all groups referred to the distribution of wins among the three outcomes in formulating reasons for their inference. Group 5 gave no reason.

At first, Group 4 claimed they had the bottle on steroids, since red won most times and that the same color came in a row. The reason, "the same color came in a row" was probably grounded in a common perception of randomness. To many people, randomness is perceived as something irregular, as a mess. If the same outcome comes in a row, in a streak, it is perceived there is more to the situation than just randomness (Sun & Wang, 2010). After a while, Sara entered Group 4 and asked them how they could be surer than 50%. In the discussion with Sara, the group proposed to do further games. Sara handed out a new worksheet of game boards to the group and they run four new games before all groups were done. Blue won the two first games, then won yellow and last red. This observation made the students change their decision. They did not get enough time to write down their explanation but in the film, they made explicit that their reason for changing to not having the bottle on steroids was because the number of wins (the mode) across the samples was rather even between the colors as no color deviated strongly from the other colors across the 12 (8+4) samples.

With small differences, all other groups followed the last line of reasoning from Group 4, regardless of whether they inferred they had the bottle on steroids or not. Their reasoning reflected hypothesis testing in that it involved aspects of expectations, based on an idea of how the modes would distribute among the three colors if the sample space had been uniform: That is, if the sample space is uniform, it is expected that the data distribution is rather even, nothing stands out.

Several groups thought they had the bottle on steroids. At the end of Lesson 2, Sara used this uncertainty to support reasoning on how to be surer. Laying the foundation of Lesson 3 Sara asked, "What can we do to be more sure?" Edvin suggested:

Edvin: We look at the statistics on our papers [worksheets].
Teacher: Yes! How can we do that?
Edvin: We can look at which color won must times i.e., in total for the whole class and then…
Teacher: So, if…I come up with, if we have six, three, two [posts 6/3/2 on the board] and then we have five, two and one [posts 5/2/1], then we compare and look [moving her hand between the two posts], is it how you are thinking?
Edvin: And then can we find the one that have won, the color that has most, the group that has most of that color, and the other colors have a bit less wins, maybe two.

From the statements, "Look at which color won most times" and "The group that has most of that color, and the other colors have a bit less wins, maybe two," it appeared Edvin seemed to focus on comparing the proportion of wins between the samples from the different bottles. What strengthens this interpretation is that Edvin belonged to Group 6, the group using proportion language to argue why they claimed they had the bottle on steroids (Table 1). Sara then asked Nathalia in Group 4 to voice the idea they discussed:

Nathalia: Test on a new paper, do a bit more [games]
Sara: Yes, we talked about doing more rounds, if that could make us more sure.

Above I discussed how Sara challenged Group 4 during the group work on how they could be surer of a decision. They developed the idea to produce more samples. By asking Nathalia, Sara wanted to make this idea explicit to the class, as a basis for Lesson 3.

4.3. THE LARGER THE VARIATION, THE GREATER THE REASON FOR REJECTING A UNIFORM SAMPLE SPACE

Lesson 3 was designed to push further students' ability to use variation in outcome frequencies across samples as data-based evidence for distinguishing a non-uniform random generator from uniform random generators and to see the need for more samples for increased certainty. To this end, the teaching should provide needs for further statistical investigations. Sara approached this need by showing the outcomes from Lesson 2 and pointing out that there is uncertainty about which bottle is on steroids because different groups believed they have that bottle (Table 1).

Table 1. Responses from students' worksheets

Group	Distribution of wins	Decision	Reasons	Degree of certainty
2	0R, 0Y, 8B	Yes	One color won all times	90%
3	4R, 2Y, 2B	Yes	One color won most times	45%
6	1R, 2Y, 5B	Yes	One color appeared so often and won five of eight times.	60%
1	3R, 2Y, 3B	No	It was even	60%
4	4R, 1Y, 3B	No	The same color appeared several times in a row	80%
5	1R, 2Y, 5B	No	Don't know	100%
7	2R, 2Y, 4B	No	All colors won at least one time	10%
8	2R, 5Y, 1B	No	All colors won at some point	50%

Sara handed out the worksheets from Lesson 2, asking the groups to discuss if they keep to their decision or want to change. Group 3 and 8 changed their decision after their discussions, which led to Group 2, 6, and 8 believing they had the bottle on steroids. Sara posted 2, 6, and 8 on the whiteboard, accompanied by the winning distributions (Figure 2). All three groups argued they had the bottle on steroids. In terms of informal hypothesis testing, the selection of the three candidates was based on the reason that the three outcomes (colors) are not likely to be equiprobable, because one outcome wins many more times compared to the other outcomes. The reason Group 3 changed their decision to not have the bottle on steroids followed a similar line of reasoning as the three other groups, with the aid of a comparison between the groups. Without any information about the samples from the other groups, Group 3 claimed, yet with a low degree of certainty, they had the bottle on steroids, due to the differences in mode frequencies across their samples. After being exposed to the samples from all groups, they saw that there were bottles that produced larger variation in mode frequencies. Based on this information, they made explicit the reason that they did not have the bottle on steroids because it is more likely one of the bottles with a larger variation in mode frequencies was on steroids.

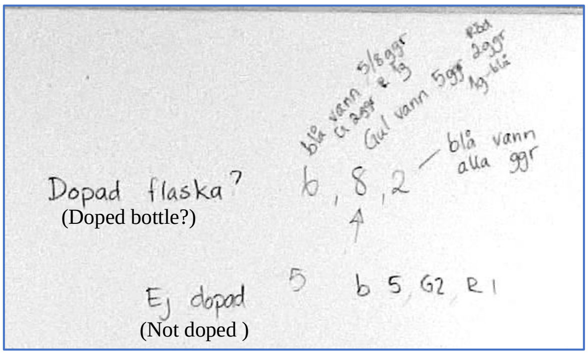

Figure 2. The four candidates of bottle on steroids

Even if the variation in mode frequencies was largest with bottle 2, it was easy for Sara to motivate the need for more samples, to gain more information for their final decision on which bottle is on steroids. But, before the students were asked to generate more samples, Sara turned to Group 5. Group 5 kept to the decision of not having the bottle on steroids. Sara knew they had a similar result as Groups 6 and 8. She used this knowledge to challenge Group 5 and to motivate further investigations on also their bottle:

Sara: Are there more groups where the same color won five times? Because I think, if we look at [Bottle] six and eight, this [pointing at "blue" for Bottle 6] won five times and this [pointing at "yellow" for Bottle 8] won five times and they think they have the bottle on steroids. Is there any other group where one color won five times?

Group 5 answered that one of their colors won five times. Sara posted 5 on the whiteboard (Figure, 2) and asked the group to explain why they think they do not have the bottle on steroids:

Teo: Because it was different colors. First it was red who won and then yellow then blue, blue and the yellow.
Sara: That it was not in a series?
Teo: Yes.
Sara: Okay, was it the same for you in Group 6 and 8 that the wins came in a series, or if did they also came irregularly?

The reasoning of Group 5 follows the perception of randomness as irregular (Sun & Wang, 2010). The wins in both Group 6 and 8 did not either come in a row and Sara asked Group 5 if they think Groups 6 and 8 also do not have the bottle on steroids. Group 5 had no answer to that. When Sara asked Group 5, they agreed on including also Bottle 5 as a candidate for further investigations.

Sara arranged the class into four groups, each with one of the four candidates. Each group also received the game board of their bottle from Lesson 2 and a blank game board with eight games. After playing eight new games, the groups should then decide if their bottle was on steroids, based on all 16 games (samples).

Sara gathered the class into a whole-class discussion after all groups had inferred if their bottle was on steroids, based on their 16 samples. With Bottle 2, blue won 15 times and the students in this group agreed on that they had the bottle on steroids. Group 6 did not reach an agreement. The wins were distributed with nine on blue, four on red, and three on yellow. Some students articulated, "They came too scattered; the colors came too scattered" as a reason that their bottle was not on steroids. In Lesson 2, the distribution of wins in Group 5 was the same as in Group 6. The same happened also in this new round of eight games. Hence, after the 16 games, Groups 5 and 6 had the same distribution of wins. The students, working with Bottle 5 agreed their bottle is not on steroids. Noted on the worksheet was, "If we compare with Bottle 2, our bottle is not on steroids." Where this reason comes from is not clear since from the video, I cannot observe that Group 5 has information of Group 2's last eight games. One way to make sense of their reasoning is that they apply what Noll and Shaughnessy (2012) refer to as additive reasoning. In other words, taking limited account of random variation, it seems as if they used the first eight games from Bottle 2 as a predictor of the second series of eight games.

The 16 games of Bottle 8 turned out as, eight on yellow, six on red and two on blue. Denise argued, "I think it is not on steroids because it was a bit even but, I am not entirely sure because it was not very even." On their worksheet the group claimed they had the bottle but was not very certain (50%). Unfortunately, it did not become clear if their decision is based on all 16 games or on only the last eight games. Nevertheless, their explanation on their worksheet, "It was quite even, but also not, about in the middle," indicates that they would have been more certain with a larger difference in the number of times each color became the mode. In relation to hypothesis testing, they may have perceived the observed deviation as likely to be the result of chance.

When the inferences from all groups had been made public, Sara asked the class to decide which bottle contained more of one color. Being aware of the 16 samples from all bottles, all students agreed on that Bottle 2 was the one on steroids. The inferences followed how Group 3 came to infer they had not the bottle on steroids in the introduction of Lesson 3. First, students noted that one of the colors win many more times than the other colors with Bottle 2. In general terms, it was noted that one outcome deviated strongly as the mode across samples from the same random generator. Second, the students noted that there was no outcome from the other bottles that stood out as the mode as strong as the blue marble did across the samples from Bottle 2. In more general terms: In the choice between two alternatives, the random generator that produces the largest difference in the distribution of the mode is chosen as non-uniform.

Lesson 3 ended with Sara asking the students if they thought any other bottle was on steroids. She asked this to further challenge the students on what would be enough difference for rejecting that a random generator is uniform, in a series of 16 samples. Some students thought Bottle 5 and 6 were bottles on steroids, while some thought not. In both cases, the variation in mode frequencies was used as an indicator, i.e., as a reason. To some students, the variation was not large enough to reject the possibility that it was only an effect of chance.

5. CONCLUDING DISCUSSION

Inferential statistics is an important topic in statistics (Sotos et al., 2007), which has proved difficult to learn (Dolor & Noll, 2015; Krishnan & Idris, 2015). To facilitate the learning of formal inferential statistics it is suggested that the learning of inferential statistics should start in the early years of schooling (Meletiou-Mavrotheris & Paparistodemou, 2015), with instructional support that allows students to engage in inferential statistics informally (Makar & Rubin, 2009; Nilsson, 2020b). Drawing on this suggestion, the overall objective of the present study has been to contribute with knowledge of learning environments, which invite students to use informal methods of hypothesis testing. A design experiment (Cobb et al., 2003) with students in Grade 5 (11–12 years old) playing Color Run (Nilsson, 2020b) constituted the context for the investigation of a hypothetical learning trajectory with the end goal of introducing students to use variations across samples as informal reasons to find out how likely it is that a concrete random generator is uniform, and the action of rejecting a random generator as uniform if the variation is large enough (cf. Zieffler et al., 2008). I discuss the contributions of the study according to the three main steps of the hypothetical learning trajectory.

The first step took form in Lesson 1. Here students were engaged in three games of Color Run (Nilsson, 2020b) with changing sample space between each game. The first step speaks to students' tendency to emphasize idiosyncratic and material reasons. It concerns how the students came to favor sample space reasoning over idiosyncratic reasoning (Makar & Rubin, 2009; Meletiou-Mavrotheris & Paparistodemou, 2015; Watson et al., 2007) when the sample space was changed between color runs (Nilsson, 2020b). The study also shows how concrete random generators can help teachers to challenge idiosyncratic and material reasons. By shaking the bottle in front of the class, the teacher made the students aware that it will not matter in which order the marbles are placed before shaking the bottle.

The student's account of sample-space reasoning challenges the equiprobability bias (Lecoutre, 1992). In line with Nilsson (2007), students considered the composition of the sample space when predicting the outcome of the three games in Lesson 1. For instance, a color that was not the most represented in the sample space won Game 2. Such an observation may support the equiprobability bias, that everything is just a matter of chance. However, this did not happen. Instead, when predicting Game 3, most students again turned to the most favorable outcome to give a reason for the color they voted for to win. As in Nilsson (2020b), it seems as if the variation in the sample space between the games turned the students' attention to the contents of the bottles and thus supported them in privileging the underlying probability distribution to give reasons for which color they expected to win.

Previous research provides a rather good picture of students' perceptions of variation (Biehler et al., 2018; Noll & Shaughnessy, 2012). We know that some students overestimate the probability of strongly deviating results (Van Dijke-Droogers et al., 2020) and that some students are willing to accept wide variation across samples (Shaughnessy & Ciancetta, 2002). The present study contributes with insights on how instruction can be designed to support students' understanding of random variation and their ability to use variation across samples to infer on how likely it is that a random generator is uniform, and for the action of rejecting a random generator as uniform. The first step of the hypothetical learning trajectory sets the foundation for this ability. The second step concerns how students came to emphasize the mode, and how the mode was distributed across samples drawn from their own bottle, to infer on if the sample space of the bottle was uniform or not. A mode distribution with five of eight wins appeared as critical in the action of inferring whether a bottle was on steroids or not. Theoretically, the probability to obtain five, or a more extreme result, in a sample of eight observations, under the null hypotheses (μ=1/3), is about 0.087. The probability to obtain six or more is about 0.019. So, considering Bottle 2, 5, 6, and 8 as candidates of a bottle on steroids shows evidence of reasonable reflections on random variation under the null hypothesis.

Understanding the role of sample size is key to understanding random variation (Lee et al., 2010). In Lesson 2, students generated a series of only eight games, each of ten steps. Eight samples were supposed to be a series long enough to obtain the bottle on steroids as one of the candidates for further investigations. It was also supposed to be a series short enough to obtain wide variations from bottles with a uniform sample space and so, include also uniform bottles as candidates for further investigations. The number of eight samples fulfilled its purpose. Several bottles were suggested as candidates to be the bottle on steroids and the teacher could use this uncertainty to engage students in discussions where students were engaged in reflecting on the need of collecting more samples.

The third step of the hypothetical learning trajectory shows how a teacher can start with hypothesis testing in an informal way, without students first having a profound understanding of relative frequencies (Ireland & Watson, 2009) and of how data distributes under the null hypothesis in the long run. As observed in Nilsson (2020b), it seems natural to the students to look at absolute frequencies and, particularly, at the mode, which is, most likely, stimulated by the idea of "winning a game." In Lessons 2 and 3, students show how they can learn to use variation in the distribution of the mode across samples to draw an inference on an unknown sample space. Their reasoning followed the logic: the larger the variation, the greater the reason for rejecting a uniform sample space. The third step adds a new dimension to the use of variations across samples compared to the second step. In the third step, the sample series from the different bottles took the role of a reference, like that of the theoretical, formal distribution under the null hypotheses. In other words, instead of comparing the outcomes only to what would be expected from a theoretical, uniform distribution (cf. Lee et al., 2010), the students compared the variations in the mode between the different bottles to strengthen their inferences. That one color won many more times than the other colors from Bottle 2, compared to how the wins were distributed from the other bottles, was used as a reason to claim Bottle 2 was on steroids, i.e., contained a non-uniform sample space of colored marbles. Hence, it seems reasonable that the learning of informal hypothesis testing takes as its starting point the variation between absolute frequencies. In the ambition to develop the hypothetical learning trajectory, research should then take a close look at how teachers can support students to expand their reasoning about variation in and between data from absolute frequencies to relative frequencies and the law of large numbers (Ireland & Watson, 2009).

To draw reasonable inferences of an underlying probability distribution, students need to evaluate and understand the role of sample size (Pratt et al., 2008). Findings from Lesson 3 add to how a teacher can help students to reflect on the role of increasing the number of samples, to infer on the underlying sample space. At first in Lesson 3, the between-bottle comparison was made based on only the first eight games, played in Lesson 2. Students noticed that there were sample series from some groups that displayed a larger variation in the mode. The teacher drew the students' attention to these sample series and to the fact that there seems to be uncertainty about which bottle was doped. The teacher could use this uncertainty to motivate students to play eight more games with these candidates. The results of the 16 samples strengthened the variation in mode between the bottles. After the 16 samples, all students were certain that Bottle 2 was on steroids.

A final remark to be made concerns that some groups marked a certainty lower than 50%. From a subjective perspective on probability, we understand a claim on certainty as a claim on probability (Borovcnik et al., 1991) and, from such a perspective, the inference that the probability is 10% (Group 7) implies the converse inference that the probability is 90% they are wrong and, consequently, that they are pretty sure they have the bottle on steroids. However, an alternative interpretation is that the students just mark how sure they are about their decision and, that they would be less sure about the alternative. Unfortunately, there was no more information on this issue since the tasks or the teacher did not challenge the student further on what a low certainty on a decision would imply for the alternative. On this account, I invite research to further investigate the meanings students ascribe to how certain they are when early engaged in drawing inferences in situations asking for hypothesis testing.

REFERENCES

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

Bakhurst, D. (2011). *The formation of reason*. John Wiley & Sons.

Ben-Zvi, D., Ainley, J., & Gravemeijer, K. (2018). Design of statistics learning environments. In D. Ben-Zvi, K. Makar & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 473–502). Springer. https://doi.org/10.1007/978-3-319-66195-7_16

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

Biehler, R., Frischemeier, D., Reading, C., & Shaughnessy, J. M. (2018). Reasoning about data. In D. Ben-Zvi, K. Makar & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 139–192). Springer.

Borovcnik, M., Bentz, H.-J., & Kapadia, R. (1991). A probabilistic perspective. In R. Kapadia & M. Borovcnik (Eds.), *Chance encounters: Probability in education* (pp. 27–71). Springer.

Brandom, R. (2000). *Articulating reasons: An introduction to inferentialism*. Harvard University Press. https://doi.org/10.4159/9780674028739

Canada, D. (2006). Elementary pre-service teachers' conceptions of variation in a probability context. *Statistics Education Research Journal*, *5*(5), 36–63. https://doi.org/10.52041/serj.v5i1.508

Cobb, P., Confrey, J., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13. https://doi.org/10.3102/0013189X032001009

Dolor, J., & Noll, J. (2015). Using guided reinvention to develop teachers' understanding of hypothesis testing concepts. *Statistics Education Research Journal*, *14*(1), 60–89. https://doi.org/10.52041/serj.v14i1.269

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Reidel Publishing.

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.

Ireland, S., & Watson, J. (2009). Building a connection between experimental and theoretical aspects of probability. *International Electronic Journal of Mathematics Education, 4*(3), 339–370. https://doi.org/10.29333/iejme/244

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. https://doi.org/10.2307/749741

Krishnan, S., & Idris, N. (2015). An overview of students' learning problems in hypothesis testing. *Jurnal Pendidikan Malaysia, 40*(2), 193–196. https://doi.org/10.17576/JPEN-2015-4002-12

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 mathematics Education* (pp. 490–525). National Council of Teachers of Mathematics.

Lecoutre, M.-P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics, 23*(6), 557–568. https://doi.org/10.1007/BF00540060

Lee, H. S., Angotti, R. L., & Tarr, J. E. (2010). Making comparisons between observes data and expected outcomes: Students' informal hypothesis testing with probability simulation tools. *Statistics Education Research Journal*, *9*(1), 68–96. https://doi.org/10.52041/serj.v9i1.388

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105. https://doi.org/10.52041/serj.v8i1.457

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.

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. https://doi.org/10.1007/s10649-014-9551-5

Mooney, E., Duni, D., VanMeenen, E., & Langrall, C. (2014). Preservice teachers' awareness of variability. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education.* Proceedings of the Ninth International Conference on Teaching Statistics, Flagstaff, Arizona. http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_C149_MOONEY.pdf?1405041831

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. https://doi.org/10.1007/s10649-006-9062-0

Nilsson, P. (2020a). 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. https://www.jstor.org/stable/10.5951/jresematheduc-2020-0167

Nilsson, P. (2020b). Students' informal hypothesis testing in a probability context with concrete random generators. *Statistics Education Research Journal, 19*(3), 53–73. https://doi.org/10.52041/serj.v19i3.56

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 AG. https://doi.org/10.1007/978-3-319-72871-1_4

Noll, J., & Shaughnessy, J. M. (2012). Aspects of students' reasoning about variation in empirical sampling distributions. *Journal for Research in Mathematics Education, 43*(5), 509–556. https://doi.org/10.5951/jresematheduc.43.5.0509

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. https://doi.org/10.52041/serj.v7i2.472

Rubin, A. (2007). Much has changed; little has changed: Revisiting the role of technology in statistics education 1992–2007. *Technology Innovations in Statistics Education, 1*(1). https://doi.org/10.5070/T511000027

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning about sampling: Trouble at the core of statistic. In *Proceedings of the Third International Conference on Teaching Statistics* (pp. 314–319).

Runesson, U. (2006). What is it possible to learn? On variation as a necessary condition for learning. *Scandinavian Journal of Educational Research, 50*(4), 397–410. https://doi.org/10.1080/00313830600823753

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.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114–145. https://doi.org/10.5951/jresematheduc.26.2.0114

Simon, M. A., & Tzur, R. (2004). Explicating the role of mathematical tasks in conceptual learning: An elaboration of the hypothetical learning trajectory. *Mathematical Thinking and Learning, 6*(2), 91–104. https://doi.org/10.4324/9780203063279

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98–113. https://doi.org/10.1016/j.edurev.2007.04.001

Sun, Y., & Wang, H. (2010). Perception of randomness: On the time of streaks. *Cognitive psychology, 61*(4), 333–342. https://doi.org/10.1016/j.cogpsych.2010.07.001

Thompson, P. W., Liu, Y., & Saldanha, L. (2007). Intricacies of statistical inference and teachers' understandings of them. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 207–231). Psychology Press.

Van Dijke-Droogers, M., Bakker, A., & Drijvers, P. (2020). Repeated sampling with a black box to make informal statistical inference accessible. *Mathematical Thinking and Learning, 22*(2), 116–138. https://doi.org/10.1080/10986065.2019.1617025

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. https://doi.org/10.1080/10986060709336812

Watson, J., & Kelly, B. (2004). Statistical variation in a chance setting: A two-year study. *Educational Studies in Mathematics, 57*(1), 121–144. https://doi.org/10.1023/B:EDUC.0000047053.96987.5f

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. https://doi.org/10.52041/serj.v7i2.469

PER NILSSON
Linnaeus University
351 95 Växjö
Sweden