Educational Studies in Mathematics (2021) 108:553–577
https://doi.org/10.1007/s10649-021-10067-8

Pre-service teachers' difficulties in reasoning about sampling variability

Omar Abu-Ghalyoun[1]

Accepted: 11 May 2021 / Published online: 10 June 2021

Abstract
Past studies have documented some pre-service teachers' (PSTs) difficulties in reasoning about sampling variability. This study adds to the body of literature by investigating the ideas that PSTs employ in reasoning about sampling variability, and by conjecturing what is behind the difficulties especially during the contextuality episodes. This issue was studied in the context of a content course on statistics and probability for elementary and middle grade PSTs at a Midwestern American university. An analysis of a PST's video and screen records of a task-based interview was guided by techniques of knowledge analysis (diSessa et al., 2016), and focused on two clear contextuality episodes that have caused disequilibrium in the PST's knowledge system. The analysis described at a moment-by-moment level the cognitive dynamics of the transitions that occurred in the PST's knowledge system and highlighted some of the difficulties that arise from the activation of less-productive knowledge elements over others. The significance of this study is in its use for the techniques of knowledge analysis from the field of cognitive science to provide a fine-grained description of PSTs' difficulties in reasoning about sampling variability that went beyond the traditional description of these difficulties as misconceptions.

Keywords Sampling variability · Pre-service teachers · Knowledge analysis · Informal inferential reasoning

1 Introduction

Recent influential policy documents, such as the Common Core State Standards (National Governors Association Center for Best Practice & Council of Chief State School Officers, 2010) and Guidelines for Assessment and Instruction in Statistics Education Report (Franklin et al., 2007), have called for dramatic changes in the statistics content included in the K-8

✉ Omar Abu-Ghalyoun
omar.ghalyoun@wmich.edu

[1] Western Michigan University, Kalamazoo, MI, USA

curriculum. In particular, students in these grades are now expected to develop Informal Inferential Reasoning (IIR) as a way of preparing them for formal concepts of inferential statistics such as confidence intervals and testing hypotheses. Ben-Zvi et al. (2007) describe IIR as the cognitive activities involved in informally making statistical inferences. Makar and Rubin (2009) identified three main features that characterize informal inferential reasoning: (i) a statement of generalization beyond the data, (ii) use of data as evidence to support this generalization, and (iii) probabilistic (non-deterministic) language that expresses some uncertainty about the generalization. Over the path from informal inference in elementary and middle grades to formal inference in high school, many important concepts will be integrated into students' understanding and therefore underpin their IIR ability. Among these concepts is sampling (Triola, 2018). Therefore, any conceptual approach to statistical inference must be built on some robust understandings of the basics of sampling bias and sampling variability (Pfannkuch et al., 2015). More recently, research has shifted the emphasis from a focus on sample size and bias to sampling variability and this, to some extent, has been facilitated by the abundance of technological tools that provide user-friendly sampling simulation tools (Ben-Zvi et al., 2018). Gil and Ben–Zvi (2010) studied Grade 6 students' ideas about random sampling and found it challenged as they generate multiple random samples from a population using *TinkerPlots*: Dynamic Data Exploration (Konold & Miller, 2015). The students seemed to be concerned that the different random samples show inconsistent results. This, to some extent, undermined students' confidence in their inferences. This is promising, because if students are familiarized with developmentally appropriate instruction in sampling variability in primary school, it will help them develop better statistical reasoning in later schooling, and appreciate the usefulness of statistics for everyday life (Makar et al., 2011). Sampling variability has been explicitly emphasized in both of the above policy documents. For example, the Common Core State Standard for Mathematics (CCSSM) recommended that grade seven students have experience with generating "multiple samples (or simulated samples) of the same size to gauge the variation in estimates or predictions" (p. 46). Sampling variability is one of the known complex and multi-faceted statistical concepts where true understanding requires learners to be aware of and competently reason with many related statistical ideas (Pfannkuch, 2008).

The complexity of sampling variability has led Pfannkuch (2008) to identify five different facets of this concept along with the ways of learners' thinking associated with each of these facets: (1) the effect of increasing the size or the number of the samples on the appearance of the expected value in the sampling distribution, (2) the effect of the sample size on the location of the expected value in the sampling distribution, (3) the shape of the sampling distribution and how it grows to become symmetric as the number of the selected samples increases, (4) the effect of the sampling method on the sampling outcomes, and (5) the overall purpose of selecting samples which is making an inference about some aspect or characteristic of the parent population.

Given this emphasis on sampling variability in the K-8 curriculum, future teachers need support to acquire sufficient content knowledge of this concept. Despite the considerable body of literature on how important it is for teachers to have a deep understanding of statistical concepts to teach statistics effectively (e.g., Burgess, 2011; Noll, 2011), very few studies have investigated PSTs' inferential reasoning processes, especially as they pertain to the concept of sampling variability that supports this reasoning.

These studies indicated that PSTs face some challenges when making informal inferences based on samples of data. In a large-scale questionnaire study, De Vetten, Schoonenboom et al. (2018) showed that most PSTs who they asked to make inferences tended to describe the sample itself instead of making conclusions that go beyond the data at hand. Therefore, they did not appear to understand that representative samples can be used to make inferences about the population. Similarly, Mooney et al. (2014) found that a substantial proportion of the PSTs they examined understood that sample distribution is likely to be different from the population distribution. De Vetten et al. (2018) also asked PSTs to determine which descriptive statistics is more appropriate as evidence-based arguments for informal inferences. The PSTs argued that global descriptive statistics can be used as evidence for the inference, but they did not recognize that local aspects of the sample distribution might not be reliable evidence. Other studies showed that PSTs tend to focus on measures of central tendency as evidence for their inferences at the expense of measures of distribution (Canada & Ciancetta, 2007); while their understanding of the mean, median, and mode is mostly procedural (Groth & Bergner, 2006; Jacobbe & Carvalho, 2011). Concerning the sampling variability, studies also indicated that PSTs face some challenges when reasoning about this concept, such as using informal terminology to describe sampling variability in preference to standard descriptions (shape, mean, etc.) (Canada & Makar, 2006), and focusing only either on the range of the distribution, or the center of the distribution or on small clusters of data rather than integrating different aspects of the distributions (Canada, 2008; Mooney et al., 2014). Watson and Callingham (2013) showed that only half of the interviewed in-service teachers in their study could conceptualize that a smaller sample has larger variability. PSTs might also not possess these understandings. However, no studies have so far been conducted on PSTs' reasoning about micro facets of the concept of sampling variability, neither on the cognitive dynamics of the transitions (e.g., the activation of knowledge elements related to some facet of this concept) that occur in their knowledge systems while they reason about these facets of this concept. The ability to argue about sampling variability and suggest new conclusions requires both a strong statistical and contextual knowledge foundation. Cobb (2007) argued that statistics is one of the hardest school subjects to teach since it is inextricably tied with contexts. In mathematics teaching, however, context might make abstract concepts accessible, but this does not necessarily require the learners to consider the context in their answers (delMas, 2004). This means that "statistics requires a different kind of thinking because data are not just numbers; they are numbers with a context" (Cobb & Moore, 1997, p. 801). The term context is usually used to refer to a wide variety of things including the data context, which is the real-world context from which the statistical problem emerged (Pfannkuch, 2011). For instance, knowledge of how the data were collected, including the design of the study and how variables were defined. Learners are also involved in other context-related activities when they interpret the data during the statistical investigation process; such as reflecting on their real-life experience and using it to support their inferences, and evaluating the beliefs that they hold about the real world (Hahn, 2014). With the significant role of the context in statistical reasoning in mind, I conjectured that contextuality episodes that appear in PSTs' reasoning about sampling variability and entail using different reasoning about the same statistical ideas across different data contexts would be the most appropriate moments to analyze to see their difficulties. In this study, I use the principles of an epistemological perspective from cognitive science called the coordination classes model (diSessa & Sherin, 1998; diSessa & Wagner, 2005; diSessa et al., 2016) to describe at a moment-by-moment level the cognitive dynamics of the transitions that occurred in the PSTs' knowledge system during some contextuality episodes. The research question that guided this study was: What knowledge elements related to sampling variability supported and constrained the PSTs' abilities to reason about sampling variability across different data contexts?

2 Theoretical perspective

The analysis of the data in this study was informed by an epistemological perspective referred to as Knowledge in Pieces (KiP) (diSessa, 1988, 1993). KiP belongs to the field of conceptual change (Vosniadou, 2013), a learning theory related to constructivism, which studies learning that is especially difficult. KiP has provided a deeper understanding of the phenomenon of "prior conceptions," a term used to describe students' intuitive, pre-instructional, everyday ideas of phenomena of the natural world and their roles in emerging competence (diSessa, 2018, p. 67). KiP tries to explain the relationship between the details of students' real-time thinking and the long-term changes in their knowledge systems, which have not been articulated well by Piagetian psychology. KiP has been used by mathematics educators to examine emerging competence in the domains of fractions (Smith, 1995), probability (Wagner, 2006), calculus (Jones, 2013), and algebra (Levin, 2018). KiP has characteristics that made it a useful candidate as a foundational approach for this study. Specifically, its fine-grained quality suits the questions of the present research, allowing a productive examination of processes of knowledge reorganization across different contexts. In doing so, it enables me to zoom in on the reasoning process and analyze the cognitive dynamics of the transitions that occur in knowledge reorganization. KiP has its roots in studies of students' reasoning about the physical world (diSessa, 1993) with many aspects of the theoretical perspective that are most apparently related to the case of reasoning about physics. That said, the program of work outlined by diSessa has implications for studying knowledge and learning processes more broadly, and thus, in using and adapting this framework to study statistical reasoning, I am contributing to the growing body of work that develops the perspective beyond its origins in physics reasoning (see diSessa et al., 2016, for a review).

2.1 Basic assumptions

One of the central ideas of the Knowledge in Pieces (KiP) perspective is that knowledge can be productively modeled as a system of diverse knowledge elements that are abstracted from experience. In particular, students' intuitive ideas are considered to be potentially productive resources—neither correct nor incorrect in and of themselves—from which more coherent and integrated knowledge systems can be developed. Knowledge systems, therefore, are considered to be made up of numerous knowledge elements each of which is not necessarily right or wrong in isolation, but is productive or not for a particular context. This view thus offers a different way of conceptualizing "misconceptions" (See Smith et al., 1993; Brown et al., 2015).

As individuals reason about new situations, they activate their prior knowledge elements. From the KiP perspective, learning is modeled largely as a process of reorganization in which, through the feedback of various kinds, the use of existing knowledge elements is improved over time to better fit the context. This is often referred to as "tuning [the knowledge system] towards expertise" (diSessa, 1993). More broadly, learning a new idea can be considered a transformation of one knowledge system into another. This process of knowledge transformation contrasts with replacement models in that there may be many common knowledge elements between the knowledge systems of beginners to a domain and experts. Yet for the experts, the same knowledge elements may be activated in more refined (and contextually appropriate) ways. Figure 1 gives a schematic representation of the above ideas related to the transformation over time from naïve to conceptually competent. It is offered as a heuristic tool

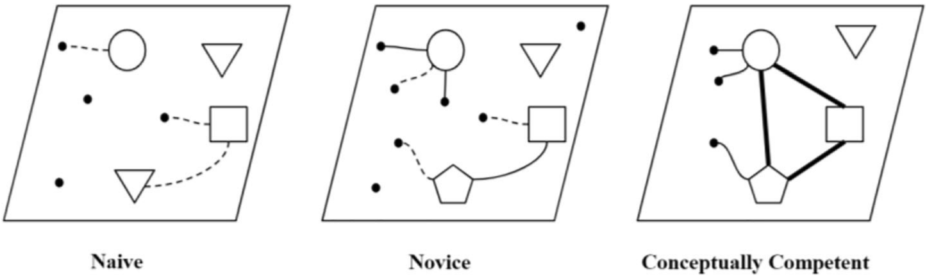

Fig. 1 Snapshots of the development of knowledge (Adapted from Reconsidering Conceptual Change: Issues in Theory and Practice, edited by Margarian Limon and Lucia Mason. Copyright ©2002 Kluwer Academic Publishers.)

in thinking about "understanding" as a phenomenon of a knowledge system (as opposed to a framework to be applied, per se). The geometric shapes in Fig. 1 are intended to communicate knowledge elements of diverse types. In the "naïve" snapshot, the connections between knowledge elements are tenuous, whereas the connections between knowledge elements in the conceptually competent snapshot are stronger. One can see that some of the same elements that existed even in the naïve knowledge system still play a role (albeit a different one) in the conceptually competent snapshot. An example clarifying the constructs of the KiP framework follows this section.

In line with Piagetian constructivism (e.g., Confrey, 1986; von Glasersfeld, 1991), the KiP perspective assumes that different contexts can be interpreted through different combinations of knowledge elements (i.e., schemas). Therefore, lack of systematicity across different contexts might prevent learners from noticing differences (contradictions), and the sensitivity of the contexts might prevent them from seeing similarities. Thus, within this view, the role of context is significant. That is, some ideas (knowledge elements) will be activated and used in some contexts, and not in others. Furthermore, as a learner's knowledge system becomes more well-organized, the contexts in which these knowledge elements get consistently activated may change. Understanding how knowledge activation and use depends on context is a core phenomenon KiP aims to capture.

2.2 More specific constructs

According to diSessa (1993), a fundamental mechanism that affects the dynamics of the knowledge systems is the *activation* and *deactivation* of existing knowledge elements. These dynamics of knowledge systems can be described in terms of *cueing priority* which is "the degree to which a particular knowledge element's transition to an active state is affected by other previously activated elements" (p. 313). A high cueing priority means that only a small additional contingent activation is still needed in order to activate the knowledge element in some context. Kapon and diSessa (2012) suggested more refined priority considerations: (a) intrinsic priority: the degree of inherent confidence in a certain knowledge element; and (b) contextual priority: the degree of confidence in the applicability of the knowledge element in some context. Within the KiP epistemological perspective, *coordination classes* (diSessa & Sherin, 1998; diSessa & Wagner, 2005; diSessa et al., 2016) are knowledge systems that model cognitive structures useful for describing particular types of concepts in physics and mathematics such as force, expected value, rate of change, etc. The coordination class model mainly addresses the difficulty that individuals face when they learn concepts that they must operate or recognize in several contexts, especially those in which they are likely to need quite different strategies to operate the concepts across contexts.

diSessa and Wagner (diSessa & Wagner, 2005; Wagner, 2010) further elaborated this model by calling the set of knowledge elements and reasoning strategies that enable the learner to recognize and apply the concept within some context the *concept projection*. The range of contexts across which the learner's concept projections are found to be applicable constitutes the *span* of the concept projections. Thus, a learner's understanding of a concept might be supported by a variety of concept projections, which in turn might consist of many knowledge elements shared by different concept projections. diSessa and Sherin (1998) called the learner's ability to recognize and apply the same concept projections across different contexts *alignment*. However, if the learner sees contrary indications of some cued concept projections associated with the same concept, the degree to which they experience any *disequilibrium* would depend on their relative confidence in each of the contrasting knowledge projections (Izsák & Jacobson, 2017). If they had high confidence in each of the contrasting knowledge projections, their experience of disequilibrium might be strong and difficult to resolve. If they had more confidence in some knowledge projections than others, their experience of disequilibrium might be resolved more easily in favor of the cued knowledge projections that provided more sense to them. Having a fully developed coordination class entails a learner's ability to (a) integrate all of the relevant information in a particular context and (b) aligning the different concept projections across the range of applicable contexts. Thus, a fully developed coordination class reflects expertise.

Using the notion of coordination classes and particularly the ideas of the concept projections, Wagner (2006) offered an alternative explanation of *knowledge transfer*. The canonical explanations in the psychological literature for transfer involve mechanisms such as structure mapping, involving a subject abstracting a context-independent knowledge structure from one context and then transferring their understanding to a new context. In contrast, Wagner's account describes the process of transfer as "incremental growth, systematization, and organization of knowledge elements that only gradually extend the span of situations in which a concept is perceived as applicable" (p. 10).

3 Methods

3.1 Data collection

Analyzing a contextuality episode in the subject's reasoning is not an easy research task because what learners do as they compare data contexts often happens in a fairly rapid way that is hard to observe. Therefore, this analysis requires very rich data that clearly exhibit details of thinking. To do so, I tried to cast a wide net across many subjects and data contexts to raise the possibility of choosing a PST who is able to provide rich and useful data for my analysis. Among seven participating PSTs, a goal in my mind as I read the transcripts of the collected data was to identify a PST who was particularly vocal and articulate, and who showed signs of active engagement in the tasks through their willingness to wrestle with the tasks and try to justify their answers. In the end, one such case rose to the top, Tanner[1], whose case I analyzed

[1] All names are pseudonyms.

in detail in this study. Tanner's major is elementary teacher education and he had not taken any statistics courses before.

For this study, I collected data from video recordings, screen recordings, and written responses during task-based clinical interviews. I also used an online questionnaire to get some background data about the participating PSTs and their current knowledge about sampling variability (assigned as homework during the first week of the semester). Each participating PST was interviewed individually for approximately 60 minutes. During the interviews, I asked the PSTs to work on three tasks, each of which involved opportunities to discuss and reason about similar facets of sampling variability but in different data contexts. These tasks were new to the PSTs, therefore, they allowed for invoking new reasoning. It is important to stress that the interviews (including the tasks and the follow-up questions) were not intended to teach new content to the PSTs but rather to provide a means for new patterns of reasoning to emerge that could be analyzed. Therefore, this study creates snapshots of PSTs' reasoning about sampling variability at one point in time rather than tracking learning over time. The interviews were conducted in week six of the semester because I expected that, based on the basic assumptions of the KiP perspective, I might be able to observe a wider range of non-normative[2] patterns of reasoning about sampling variability at this early time of the semester.

The study of sampling variability typically focuses on taking repeated samples from a population and comparing sample statistics, (such as the sample mean or the sample proportion). Therefore, the design of the tasks in this study was informed by the *black box* sampling approach, which has been shown in the literature to support the accessibility of the concept of sampling variability (e.g., van Dijke–Droogers et al., 2020). In this approach, learners make inferences about the content of a box filled with marbles or beans by collecting repeated samples and comparing results with different sizes and different numbers of repetitions. Given the small number and size of the samples that can be generated in the black box sampling approach, the outcomes might not clearly show the features of the data and therefore not help learners access the concept of sampling variability. Purposefully designed computer-based simulations, which can generate a huge number of simulated samples, along with physical simulations might help learners access this concept (Shaughnessy, 2007; Browning et al., 2014). Therefore, one of the interview tasks in this study was designed based on the dynamic data visualization software TinkerPlots. Below is a brief description of the three tasks used during the interview. More details about these tasks will be provided later in the data analysis section.

The Bean Task The goal of this task was to estimate the percentage of red beans in a container that held red and white beans as shown in Fig. 2. The actual percent of the red beans in the container was set at 20%. PSTs were asked to scoop up samples using two scoops of different sizes, develop sampling distributions, and then use their displays to draw inferences about the number of the red beans in the container (the population). The samples vary in size, but they range around 12 beans per small scoop and 35 per large scoop. They also discussed the effect of using the larger scoop instead of the small one on the shape of the sampling distribution and the confidence of their inferences using each size of the scoop.

[2] I use "normative reasoning" to indicate reasoning that is statistically accurate and/or appropriate for the context and "non-normative reasoning" to indicate reasoning that is either statistically inaccurate or not applicable to the context at hand.

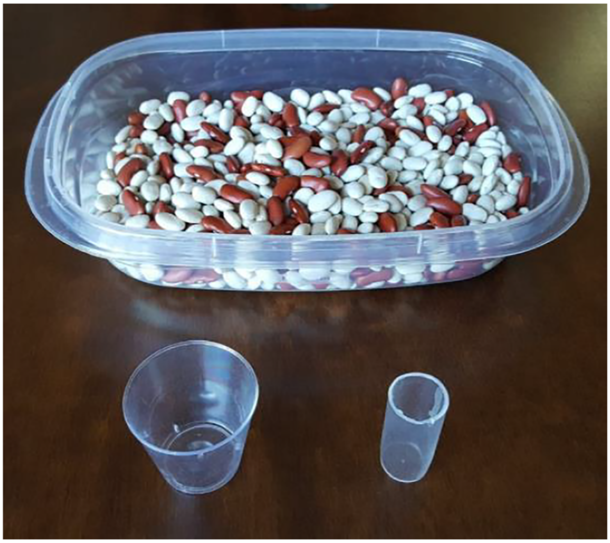

Fig. 2 The population and the scoops used to select the samples in the Bean Task

The Voting Task This interview task included the use of a pre-designed sampler that simulated collecting multiple samples of the same size (n=15) from California state voters who have been asked whether they would vote for or against some proposition. The sampler was designed using TinkerPlots. As shown in Fig. 3, the sampler had a graphical display for the immediate sample (the upper box), and a graphical display for the history of the collected samples (the lower two boxes). The sampler was set to 63% of California state voters who vote for some proposition. The clustering around this value can be found in the sampling history

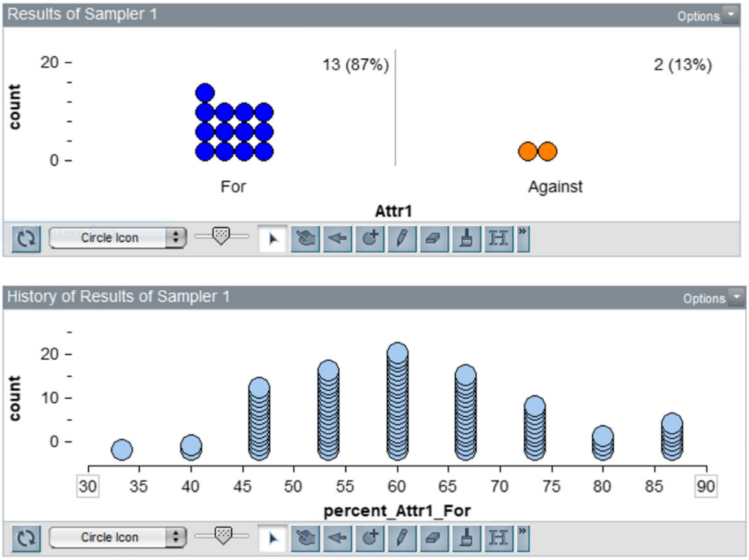

Fig. 3 TinkerPlots Sampler used in the in-class Voting Task

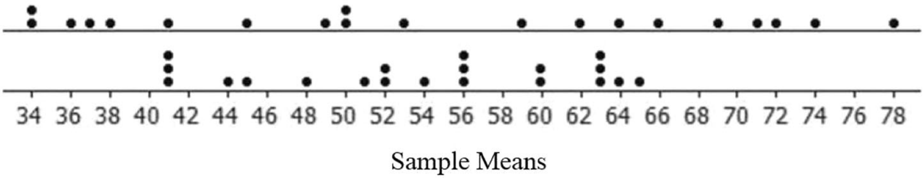

Fig. 4 The sampling distributions provided in the Gym Task

display (sampling distribution) as we generated more samples. Setting the sampler to cluster data around a pre-determined value allowed for focused interview questions.

The Gym Task Similar to the previous tasks, the Gym Task involved a new data context and required reasoning about the notion of sampling variability. In this new data context, the question "What is the typical time spent at the gym?" was investigated by selecting many samples from a population of 800 gym members. Two different sampling distributions (dot plots) of sample means calculated from random samples of the population were given as shown in Fig. 4.

These two sampling distributions are different in their spread and ranges. The discussion questions focused on comparing the shapes of these two sampling distributions and what might have caused these differences including the size of the selected samples in each case. Also, the reliability of each sampling distribution for making inferences about the parent population was discussed. Through each of the above three tasks, PSTs were encouraged to make multiple arguments and externalize their thinking about the effect of the sample size and the number of the selected samples on the variability of the sampling outcomes and on the confidence of the drawn inferences.

3.2 Analytical approach

As mentioned before, I chose the case of Tanner to analyze with respect to the in-the-moment contextuality of his reasoning and the dynamics of his reasoning about sampling variability. I developed and used four analytical strands informed by a methodological approach called knowledge analysis (diSessa et al., 2016). This methodological approach, which is devoted entirely to the methodological issues related to Knowledge in Pieces, is aligned with the epistemological assumptions of Knowledge in Pieces. The data analysis started with adding observational comments (descriptive interpretations) to each line in the transcript of Tanner's interview about (i) behavioral indicators such as what Tanner did as he spoke, (ii) affective indicators such as puzzlement or confidence, bolstered by observations related to fluidity of speech (e.g., pauses), (iii) what the speaker was referring to, or what appeared on the screen or the board at that moment, and (iv) any changes in strategies used by Tanner. I read the whole transcript, including the observational comments, carefully looking for any episodes that might illustrate the phenomenon of using different reasoning about the same statistical ideas across different data contexts. This resulted in identifying two of the clear contextuality examples (episodes). I described in detail the sequence of the reasoning that led to each instance of contextuality and the changes in the reasoning process across the different data contexts. After that, I identified some of the constructs associated with the KiP theoretical framework (concept projection, knowledge elements, alignment, etc.) in Tanner's words, then used these constructs to characterize in fine-grained detail the difficulties using the theoretical machinery of Knowledge in Pieces. To increase the reliability of the identification of the knowledge elements in this data analysis, I used criteria developed by Kapon and diSessa (2012) specifically for this purpose. First, I looked for the frequency (stability) of the expression in Tanner's reasoning. I considered the expressions that occurred more than once in the data corpus. Second, I compared the appeared knowledge elements with similar knowledge elements identified in the literature. Even if the knowledge element does not come from the literature, but has a similar form to some well-defined knowledge element in the literature, it can lend credibility to the analysis. I found some similarities between Tanner's expressions about the concept of sampling variability and the knowledge elements related to the "law of large numbers" in probability identified in Wagner (2006). For example, the knowledge element "the larger you sample, the more you get closer to your expected value" (p. 9) identified in Wagner's study has increased the reliability of the identification of the knowledge element "sampling variability decreases as the size of the samples increases" in the current study.

4 Findings

The analysis reported in this section offers an explanation of the difficulties experienced by Tanner as he reasoned with varying degrees of success about two facets of the notion of sampling variability across different data contexts. These facets are: (i) the effect of the size of the samples on the sampling variability, and (ii) the relationship between sampling variability and inferences. The analysis uses snapshots of Tanner's processes of reasoning to provide a refined understanding of what knowledge elements he was activating in each data context and offers conjectures for what is behind these differences (e.g., What was he attending to in each data context and how did that impact what knowledge elements were activated in each data context?).

4.1 Contextuality episode one: the effect of the size of the samples on the sampling variability

The interview began with the Bean Task. In this task, Tanner was asked to select 10 samples using the small scoop and calculate the percentage of the red beans in each of them. The percentages that he got were: 23%, 14%, 33%, 0%, 23%, 26%, 27%, 36, 9%, and 0% as shown in Fig. 5. The samples vary in size, but they range around 13 beans per scoop. Tanner did not return the beans and tried to randomize the selection by looking away while scooping.

To probe his understanding of the idea that sampling variability will decrease as the sample size increases, he was asked about the possible effect of using the larger scoop instead of the small one to select these 10 samples as shown in the following dialogue[3]:

O: Assume we would like to use this [*large*] scoop instead of this [*small scoop*] and do the same thing [*selecting10 samples*]. What will happen? What would you expect to get

[3] In this analysis, I refer to myself in the transcription as "O" or "I." The transcription conventions used are the following: (a) "[…]" for a break in the speech, typically including a pause, when restart or new direction; (b) "[*Italic*]" for interpretive and informal commentary, including references to particular displays; (c) pictures of the sampling outcomes or the screens are embedded in the text; and (d) no deletions have been made from the transcript segments provided.

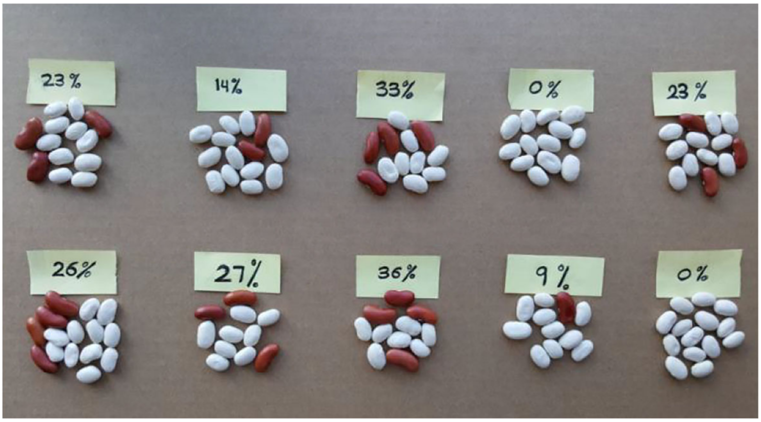

Fig. 5 Recreation of the ten samples selected by Tanner using the small scoop.

> using this large scoop?
>
> T: I would expect not as much variability between. So, I probably wouldn't expect to have 0%'s just because it's [*the larger scoop*] bigger so it [*the larger scoop*] gets more of the population, a bigger sample. So I would expect it [*the percent of the red beans in the larger scoop*] to be closer to like the middle of what we found there.

Tanner noticed that two of the samples that he got using the small scoop had 0% red beans. Because it is highly unlikely to get such a sample using the larger scoop, he inferred that large samples (scoops) will decrease the sampling variability. This noticing has led him to activate the knowledge element that "increasing the size of the samples decreases sampling variability." Recall that *concept projection* stands for the reasoning strategies that enable the learner to recognize and apply the *knowledge element* within some data context. Therefore, activating this knowledge element and applying it productively in this data context indicates the construction of some concept projection associated with the effect of the sample size on sampling variability. In view of the KiP perspective, this concept projection is among the collection of concept projections that Tanner needs to construct (maybe in a wide *span* of data contexts) and also *align* in order to construct expertise (coordination class) about this facet of sampling variability. Before we discuss the alignment across different data contexts, let us see what this concept projection entails about the relationship between decreasing sampling variability and the range of the sampling distribution. To answer this question, I first asked what he meant by "sampling variability." As shown in the following excerpt, Tanner answered the question that I was planning to ask next which is about the effect of the sample size on the range in particular.

> O: What do you mean by variability?
>
> T: Like this [*Tanner pointed to the data values on the sampling distribution, Fig. 6*] ranges all the way from 0 to 36%. I would expect it [*the data values*] to be like not as much of a range in between there [*Tanner swept back and forth between 0% and 36%*].

That he mentioned the "range" in his last answer provides evidence that his concept projection entails an awareness of the effect of increasing the size of the samples on the range in this data context. Table 1 shows the knowledge elements that Tanner has applied to the Bean Task.

From a statistics point of view, some nested relationship can be determined between the above knowledge elements. Figure 7 shows these nested relationships.

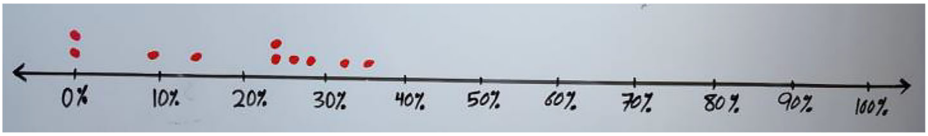

Fig. 6 The sampling distribution of the 10 samples selected by Tanner using the small scoop

So far, Tanner has activated and applied a knowledge element in the Bean Task, which is that increasing the size of the bean samples will decrease the range of the sampling distribution. The question now becomes, can Tanner apply this knowledge element productively in a different data context? The answer came shortly during the Gym Task when Tanner was asked what might have caused the difference in spread between the two given sampling distributions (Fig. 4)? Based on his answer in the Bean Task, he was expected to mention the effect of the size of the samples as one of the potential reasons. Surprisingly, he did not apply the previous knowledge element in this data context and argued that the sampling method might have caused the difference between the two sampling distributions. It was assumed in the Gym Task that each of the two given distributions represents the samples selected by a different person.

> O: What do you think might have caused the difference between these two distributions?
> T: I think it could be the method of sampling, like how they asked people. One of them might have asked their friends or something like that. Which like finding more of an accurate way to represent which is kind of what I was saying with the first one [*the upper sampling distribution in Fig. 4*], that kind of seems more like they did more of a representation of the whole population rather than the second one since it's more pushed together. It seems that the same type of people was asked.

In this response, Tanner has activated a new knowledge element associated with the effect of the sampling method on the sampling outcomes. This knowledge element is relevant because statistically speaking, the sampling method is one of the factors that influence the shape of the sampling distribution. Therefore, we cannot consider this answer as evidence that his understanding of this facet of sampling variability is emerging. Tanner had four sessions (100 minutes each) of class instruction across which sampling methods were discussed. I conjecture that recognizing the relevance of the sampling methods in this task exhibits the existence of some knowledge element associated with the sampling methods and their influence on the sampling outcomes. Within his knowledge system, it seems that the knowledge element associated with sampling methods has a high *intrinsic priority*; therefore, it is the reasoning that is most readily available to Tanner. Probing his understanding of the relationship between the size of the samples and sampling variability requires a judicious strategy that draws his

Table 1 Knowledge elements applied to the Bean Task

Knowledge element	Data context	Normative/non-normative
1. Sampling variability decreases as the size of the samples increases.	The Bean Task	Normative
2. The range of the sampling distribution shrinks as the size of the samples increases.	The Bean Task	Normative
3. The lower bound of the range increases as the size of the samples increases.	The Bean Task	Non-normative

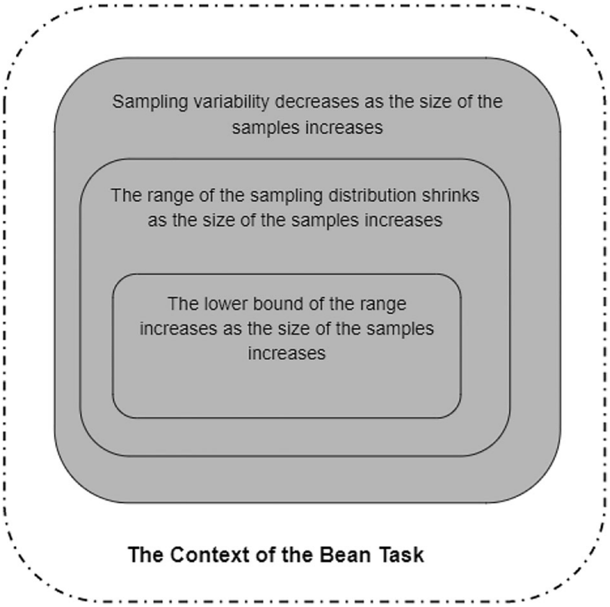
Sampling variability decreases as the size of the samples increases
The range of the sampling distribution shrinks as the size of the samples increases
The lower bound of the range increases as the size of the samples increases
The Context of the Bean Task

Fig. 7 A representation of the relationships between the knowledge elements applied to the Bean Task

attention to this relationship without telling the answer. Therefore, I pushed him to reconsider his reasoning by drawing his attention to the fact that the sizes of the samples from the two cases in the Gym Task were not the same. This cued him to think about the size of the samples as an influential factor in the shape of the distribution.

> O: What if I tell you that each of the two guys used a different sample size? Does that matter?
> T: Yes, I think it does matter because like if the second person selected small samples, then that's why all their stuff may be close together. If you have more, a bigger sample, you represent more of the population. So if you sample like ten people and the other person does 50, the other person is going to have a more variety, therefore, more accurate representation of the large population. So it definitely does affect it.

Surprisingly, the above answer did not match his previous answer in the Bean Task because this time he argued that the sampling distribution with the scattered data better represents the diversity that naturally exists in the large population. He also argued that these scattered sampling outcomes resulted from large samples as opposed to the small samples that led to compact sampling outcomes. In view of the KiP perspective, Tanner has inferred a new knowledge element associated with the relationship between the size of the samples and sampling variability and applied it in the Gym Task. His later response indicates that the context-sensitivity of his answer in the Bean Task might have impeded him from seeing similarities between the two questions. In other words, they were the two samples with 0% red beans in the Bean Task that have helped him infer the knowledge element and apply it in the data context of the Bean Task. However, the Gym data context had no similar sample with 0% that might have helped him recognize the old knowledge element and apply it in this data context. A successful learning process is always reflected in the alignment of the concept projections associated with the same notion across different contexts. This apparently has not been accomplished yet in the case of Tanner because he did not notice any discrepancy between his answers across the two tasks. To induce him to address this apparent contradiction, I pointed to the 10 bean samples (still placed on the table) and asked him to summarize his previous answer about the effect of increasing the size of the bean samples. He reaffirmed the original answer that the range will shrink as the size of the samples grows but with a clear hesitation.

> T: I guess the range could grow. Well, actually, no, I think it will shrink because there's less of a chance of you getting 0% of red beans with that cup than there is with that one. But then there's also a bigger chance of you getting more than like four beans or 36% [*Tanner pointed to the sample with four red beans which has the maximum percentage among the ten samples*] of red beans. So I could see it going either way. I guess it would shrink just because I don't think you'd have 0%, or you'd have less 0%s.

He first said the range of the bean samples could grow then changed his mind and reaffirmed the original answer and justified it using the same reasoning again based on the 0% sample. Tanner tried to apply the new concept projection that he constructed in the Gym Task by justifying the increase of the range of bean distribution. As he was trying to justify the increase of the range, Tanner thought that it is more likely to get red beans in the larger scoop (sample) therefore the percentage of the red bean in the larger samples is more likely to be higher; therefore, the upper bound of the range will increase. Tanner was talking about the number of red beans in each sample instead of the percentage of the red beans. What Tanner missed here is that the larger samples will also contain a larger number of white beans. Although this is a proportional reasoning idea, it is of special importance to sampling variability because estimating population proportion is one of the common "point estimates" in statistics. However, Tanner then changed his mind and ended with the conclusion that the range will shrink. It was not clear why he finally concluded that the range would shrink although the upper bound of the range would grow. Perhaps he thought that the increase of the lower bound would exceed the increase of the upper bound therefore the range would shrink. This hesitation indicated some effort to align the two concept projections. Also, the reaffirmation of the previous answer might indicate that this knowledge element still has high intrinsic priority even after activating a new knowledge element in the Gym Task. Tanner might have not seen the contradiction between these two knowledge elements clearly because he did not see a clear similarity between the two data contexts. Table 2 shows the knowledge elements that Tanner has activated and applied during the Gym Task.

Figure 8 illustrates the relationships between the knowledge elements applied during each of the Bean and the Gym Tasks. So far, Tanner had not noticed any discrepancy; therefore, the figure does not show any bond or discrepancy between the two data contexts.

When I asked why he thought that increasing the size of the samples would shrink the range in the Bean Task but expand it in the Gym Task he seemed not sure what to say. He just repeated his reasoning when responding to the Gym Task and emphasized that the range should grow in that case. It seems that my question was said with some surprise in a way that indicated to him that these two tasks are the same and therefore they should have the same answer. As a result, he experienced disequilibrium in his knowledge system.

Table 2 Knowledge elements applied to the Gym Task

Knowledge element	Data context	Normative/non-normative
1. Sampling variability increases as the size of the samples increases.	The Gym Task	Non-normative
2. The range of the sampling distribution grows as the sampling method becomes more random.	The Gym Task	Non-normative
3. The range of the sampling distribution grows as the size of the samples increases.	The Gym Task	Non-normative
4. Both the lower and upper bounds of the range increases as the size of the samples increases.	The Gym Task	Non-normative

> O: Why you think that increasing the size of the samples here [*the Gym Task*] will not make the range shrink?
>
> T: Interesting. Because it's more representative of the data. So, I guess it would grow because you're asking more or you're using more people in your sample, so you're getting more variety in your answers. So, I guess it's [*selecting larger samples*] more representative overall because you're asking more people. So, I guess the range would grow for this [*the Gym Task*].

As shown in Fig. 9, Tanner's knowledge system still contains all of the knowledge elements and concept projections that he has constructed so far, although he became aware that there should be something wrong in one of his answers because he got the impression from the tone of my question that they should be the same.

So far, he does not know even which of his answers should be changed because each of them is based on a knowledge element that has a high contextual priority in his knowledge system. Recall that these two contrasting knowledge elements are "increasing the size of the samples would shrink the range in the Bean Task but expand it in the Gym Task." In the following interview segment, I tried to help him resolve this disequilibrium by pointing to one of the 0% in bean samples, then asking him if large samples usually produce outliers. I was hoping this would help him see the knowledge element that he has activated in the Bean Task

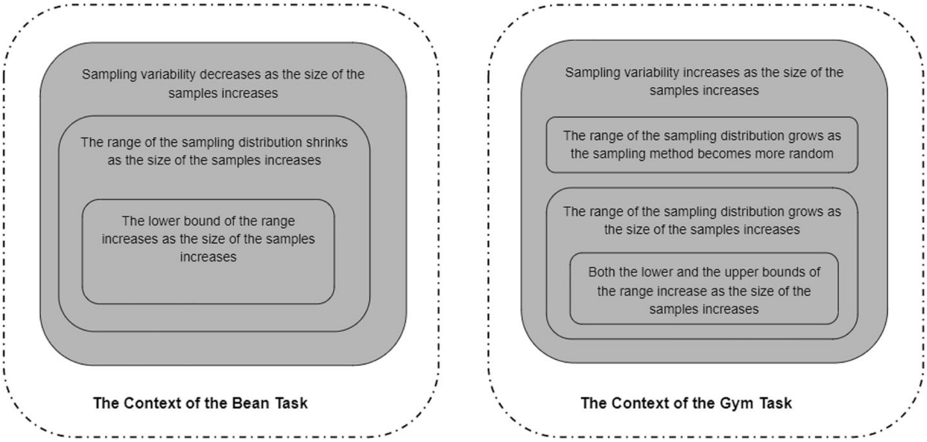
Sampling variability decreases as the size of the samples increases
The range of the sampling distribution shrinks as the size of the samples increases
The lower bound of the range increases as the size of the samples increases
The Context of the Bean Task
Sampling variability increases as the size of the samples increases
The range of the sampling distribution grows as the sampling method becomes more random
The range of the sampling distribution grows as the size of the samples increases
Both the lower and the upper bounds of the range increase as the size of the samples increases
The Context of the Gym Task

Fig. 8 A representation of the knowledge elements about the relationship between the size of the samples and sampling variability applied to the Bean and the Gym Tasks

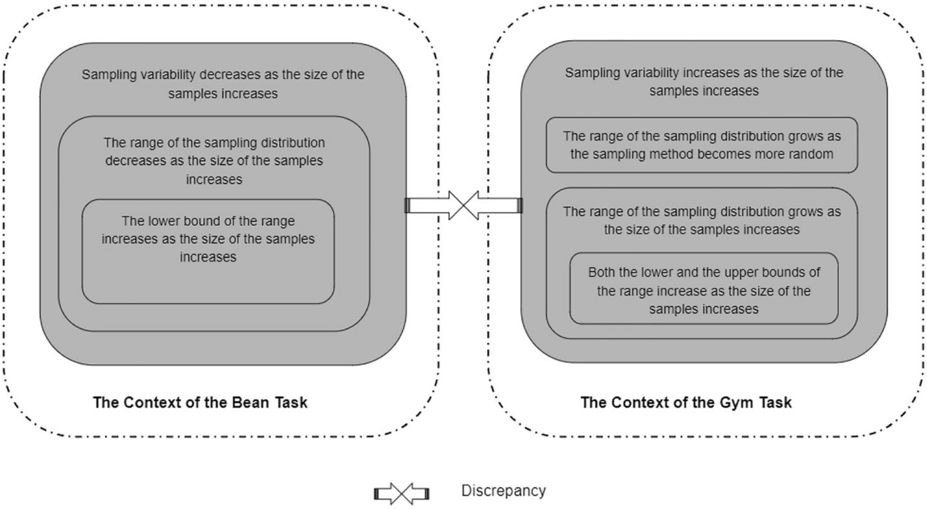
Sampling variability decreases as the size of the samples increases
The range of the sampling distribution decreases as the size of the samples increases
The lower bound of the range increases as the size of the samples increases
The Context of the Bean Task
Sampling variability increases as the size of the samples increases
The range of the sampling distribution grows as the sampling method becomes more random
The range of the sampling distribution grows as the size of the samples increases
Both the lower and the upper bounds of the range increase as the size of the samples increases
The Context of the Gym Task
Discrepancy

Fig. 9 A representation of the knowledge elements applied to each context after cueing Tanner's attention to the equivalence of the Bean and the Gym tasks

as relevant to the Gym Task. My question made him adjust his definition of "less sampling variability" which he provided earlier in the Bean Task. Recall that he described "less sampling variability" in the Bean Task as a smaller range of the sampling distribution.

> O: Do you think large samples produce outliers [*I pointed to the 0% sample*]?
> T: I think my original point [*in the Bean Task*] was that I would get more similar values. Instead of the range being from 0 to 36, I'd get more similar, like a cluster of results with just a few outliers rather than having this big of a range. So now I'm going back to my original answer to that sampling variability would be smaller, I think.

The above response indicates that he has tried to resolve the disequilibrium by applying the bean's knowledge element in the gym's context. Because the gym's knowledge element has a high contextual priority, he had to carefully justify any change in its associated knowledge element, namely, "sampling variability increases as the sizes of the samples increase." Recall that contextual priority refers to the degree of confidence in the applicability of the concept projection in some context. The above response indicates that Tanner has deactivated the knowledge element "sampling variability increases as the sizes of the samples increases." He also *inferred* a new knowledge element that "decreasing sampling variability—as a result of increasing the sizes of the samples—wouldn't necessarily mean decreasing the range but would rather mean getting more data values that would be closer to each other, not excluding the possibility of getting outliers that cause the range to increase." I used the word "inferred" to underscore that this knowledge element was created in this context rather than being an old one that he just activated. Statistically, this answer does not represent normative reasoning but it was convincing for him.

Tanner thought that he had removed the discrepancy by this answer, but I think the contrasting concept projection associated with the knowledge element that large scoops (samples) are not expected to scoop up outliers remained active in his knowledge system. Therefore, the lower bound of the range would increase as the sizes of the samples increased. Tanner had never negated this idea in any of his responses, therefore, I think this concept

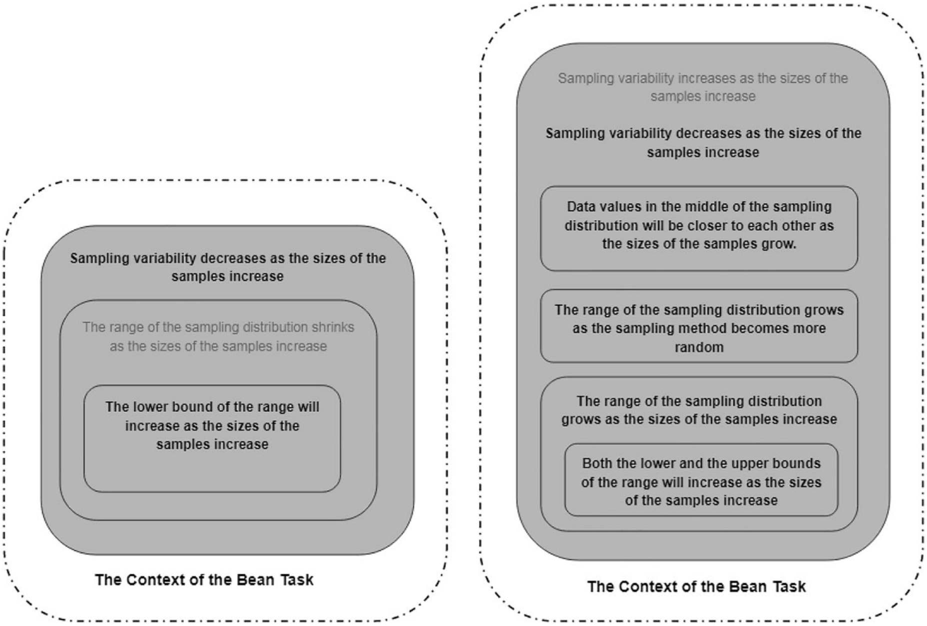
Sampling variability decreases as the sizes of the samples increase
The range of the sampling distribution shrinks as the sizes of the samples increase
The lower bound of the range will increase as the sizes of the samples increase
The Context of the Bean Task
Sampling variability increases as the sizes of the samples increase
Sampling variability decreases as the sizes of the samples increase
Data values in the middle of the sampling distribution will be closer to each other as the sizes of the samples grow.
The range of the sampling distribution grows as the sampling method becomes more random
The range of the sampling distribution grows as the sizes of the samples increase
Both the lower and the upper bounds of the range will increase as the sizes of the samples increase
The Context of the Bean Task

Fig. 10 A representation of the knowledge elements after resolving the disequilibrium about the relationship between the size of the samples and sampling variability

projection was still active in his knowledge system, but perhaps he could not recognize its relevance in the gym's data context. As shown in Fig. 10, Tanner has changed the knowledge element that he constructed while answering the Gym Task. Knowledge elements highlighted in bold type were assumed to be actively used in the solution to the task.

4.2 Contextuality episode two: the relationship between sampling variability and inferences

Another sampling variability facet that Tanner wrestled to align his concept projections with was the relationship between sampling variability and the confidence in the drawn inferences. In the following, I use snapshots of Tanner's responses during the interview to describe the constructs in his reasoning and clarify the changes in his knowledge system across the bean and gym's data contexts. Early during the interview, I asked Tanner whether he would like the sampling outcomes to be similar or different if he makes conclusions (inferences) about the percentage of the red beans in the population. His answer was:

> T: I guess similar so that you can see, like, without these being similar, you wouldn't have that cluster, so you wouldn't be able to like see exactly where it's going to be.

This response suggests that Tanner would like the sampling variability to be small in order to see clusters that help with making inferences. From a KiP point of view, this response indicated the activation of the two nested knowledge elements in his knowledge system, which are: less sampling variability is desirable for making inferences about the parent population and within this knowledge element, there was another knowledge element, which is: clusters are helpful because they help with determining the expected value as illustrated in

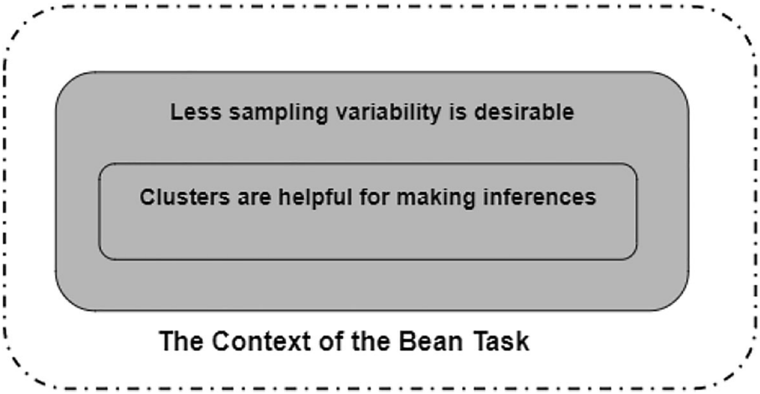
Less sampling variability is desirable
Clusters are helpful for making inferences
The Context of the Bean Task

Fig. 11 A representation of the knowledge elements associated with the inverse relationship between sampling variability and confidence in the drawn inferences applied to the data context of the Bean Task

Fig. 11. He also constructed a concept projection associated with these knowledge elements because he managed to apply and justify them productively in this context.

Statistically, this answer seems sufficient and appropriate for this question. However, from the KiP perspective, what is still needed is to affirm that Tanner can apply this knowledge element productively across different data contexts. In a later episode during the interview, Tanner was asked to answer a similar question in a different data context. That is, he was asked which of the sampling distributions in the Gym Task (Fig. 4) he preferred if he made a conclusion about the parent population.

> T: I guess the second one [*Tanner is referring to the lower sampling distribution in the Gym Task which is more compact*], just because it has less sampling variability and like there are those small clusters. It will be easier for me to find right away its mean. But I guess making it easier to find the mean doesn't really make it more accurate. I don't know. I'd still say the second one.

Neither of the sampling distributions in the Gym Task contained a clear cluster but one of them notably had a smaller range. As he was trying to apply the previous knowledge element in this new data context, Tanner noticed that one of the sampling distributions has less sampling variability but without a clear data cluster. This, however, has caused some trouble for him because he has a knowledge element in his knowledge system with high *intrinsic priority* associated with the importance of seeing a clear cluster but he cannot apply it here. He tried to justify why he preferred the compact sampling distribution for making inferences, even without having a clear cluster, by saying that it will be easier for him to find its mean. This indicates that he has activated a new knowledge element associated with finding the mean but he has noticed that this knowledge element was not relevant to this context; therefore, he has immediately deactivated it. The statement "making it easier to find the mean doesn't really make it more accurate" that follows in his response indicates the deactivation of the new knowledge element. The hesitation in Tanner's answer may have indicated that he had experienced some disequilibrium caused by, on the one hand, his recognition of the relevance of the previous knowledge element and, on the other hand, his inability to apply this knowledge element in the present context smoothly. This disequilibrium also indicated that his first concept projection about the inverse relationship between sampling variability and the confidence in the inference bean's data context was narrow in a way that does not consider the range but rather focuses on the cluster. It seems that he thought that a desirable sampling variability meant a clear cluster, like the beans sampling distribution, without any consideration for the range. With the deactivation of the new knowledge element, Tanner still had the same concept projections in his knowledge system compared to what he had constructed in the bean's data context. Figure 12 illustrates the knowledge elements associated with the inverse relationship between sampling variability and the confidence in the drawn inferences after the first response in the gym's context. The deactivated knowledge element is illustrated in gray.

To test the robustness of the last concept projection, I asked him "why do you prefer the second sampling distribution while it has no clear cluster?" He answered as follows:

> T: Because the data is more like, I don't know. Maybe I would do the first one [*the spread out sampling distribution in the Gym Task*] actually because the range makes me verify that there was a good method of sampling done because there is such a range of means, rather than the second one is kind of just all pushed together. So, I guess the first one, even though it might not visually, like help me make the assumption of what the mean would be. But I feel like it is, like since it such a variety, you can kind of find the mean of all of those and it will be your answer even though those like outliers will affect them.

In the above response, Tanner has drastically changed his previous answer and argued that more sampling variability is better for making inferences because the shape of the resulting sampling distribution will be more similar to the parent population. He also thought that although a spread-out sampling distribution does not visually help with estimating the expected value, its mean is a good estimation for the expected value. One possible explanation is that Tanner thought that a mean is a data point within the population, and therefore gathering more means should result in a distribution that is strongly similar to the population. In the above answer, Tanner has activated a new knowledge element that "more sampling variability

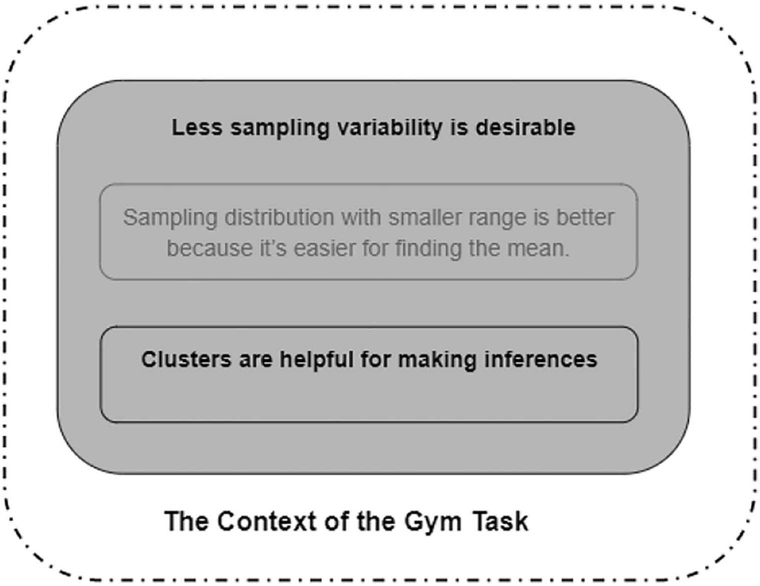
Less sampling variability is desirable
Sampling distribution with smaller range is better because it's easier for finding the mean.
Clusters are helpful for making inferences
The Context of the Gym Task

Fig. 12 A representation of the knowledge elements associated with the inverse relationship between sampling variability and confidence in the drawn inferences after the first response in the gym's data context

is better" and deactivated the contrasting one in the same data context (the Gym Task) without noticing any discrepancy with the concept projection that he has constructed in the bean's data context. By justifying the use of the mean as an alternative of the clusters for estimating the expected value, Tanner seemed to raise the contextual priority of the new knowledge element in the gym's context. The knowledge element associated with the importance of the clusters remained active in his knowledge system as illustrated in Fig. 13.

Before the follow-up question, "why do you prefer the second sampling distribution while it has no clear cluster?" Tanner seemed to be competent with the relationship between sampling variability and making inferences because he managed to apply a knowledge element associated with it in two different data contexts. His subsequent account, however, seemed to be deficient because it is not statistically accurate. My question might have shifted his attention and indicated to him that there should be something wrong with his answer. diSessa (1996) found that learners might provide both normative and non-normative reasoning in response to the interviewer's deliberate "shifts in attention" to different aspects of some concept. If Tanner had a robust concept projection, which is not the case, then he should have not changed it in response to my shifts in attention strategy. So far, the interview dialogue has not returned to the Bean Task in which Tanner preferred less sampling variability in order to

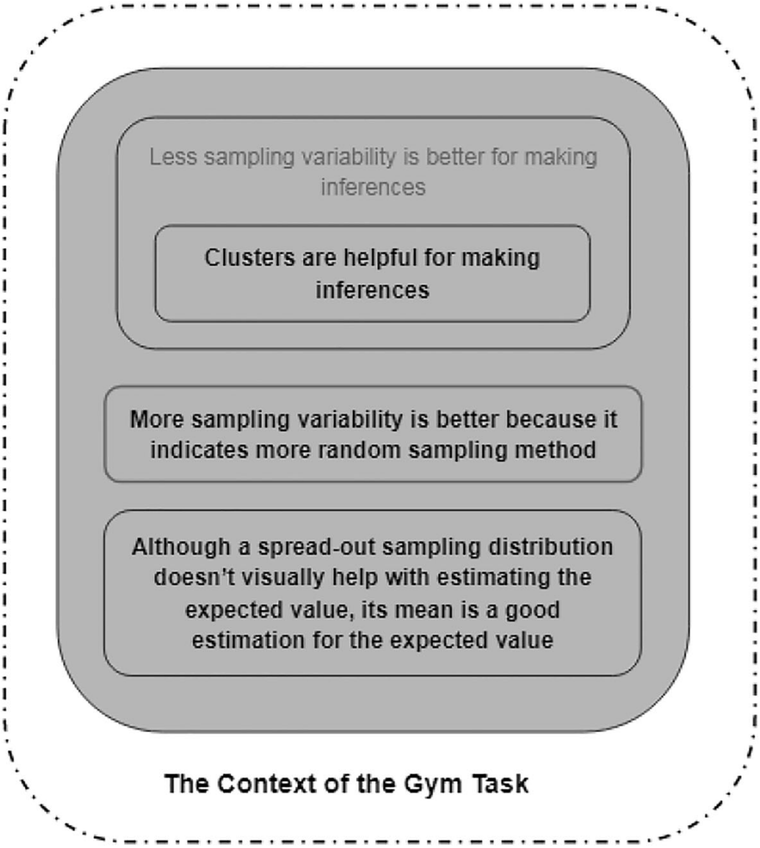
Less sampling variability is better for making inferences
Clusters are helpful for making inferences
More sampling variability is better because it indicates more random sampling method
Although a spread-out sampling distribution doesn't visually help with estimating the expected value, its mean is a good estimation for the expected value
The Context of the Gym Task

Fig. 13 A representation of the knowledge elements associated with the inverse relationship between sampling variability and confidence in the drawn inferences after the second response in the gym's data context

make inferences. Therefore, Tanner has not yet noticed any discrepancy between the two concept projections constructed across two of the interview tasks. In the following question, I reminded him of his answer in the Bean Task.

> O: So, you said you prefer the upper one [*The upper sampling distribution in the gym context, Fig. 4*]. But you had a previous answer in the Bean Task in which you said that I would like the outcomes to be similar. The outcomes here […] are not similar!
> T: Yeah. I know. I keep going back and forth because I see good and bad things in both of them. I don't know. I'm kind of leaning towards the first one *[Tanner refers to a set of data with more variability]* being so spread out, I like more, because it seems like there was a better sampling method, just because of the variety of results and that's what it would be like, I think, if you sampled the whole population, if you were able to like ask the whole population, I think it would be so spread out.

Reminding Tanner of his previous answer made him aware of the existence of some discrepancy. Saying "I don't know" indicates that he has experienced some disequilibrium in his knowledge system for a moment. Yet, this did not push him to rethink his recent inaccurate answer. Instead, he preferred it over his answer in the Bean Task which was more accurate statistically because less sampling variability always indicates a better sampling method. By leaning toward his later answer in the Gym Task, Tanner seemed to have deactivated the knowledge element that more sampling variability is better for making inferences. This, however, resulted in lowering the *contextual priority* of the concept projection that he has constructed in the Bean Task.

5 Discussion and implications

This study provided a partial replication of prior research findings showing that PSTs face some difficulties when reasoning about sampling variability especially the relationship between the center and the spread of the data (Canada, 2008; Mooney et al., 2014). This study also adds to the body of literature by highlighting PSTs' difficulties related to two facets of the concept of sampling variability: (1) the effect of the size of the samples on the variability of the sampling outcomes, and (2) the effect of the size of the samples on the confidence of the drawn inferences based on these samples. Tanner has experienced a disequilibrium in his knowledge system and faced alignment difficulties with the above two facets of the concept of sampling variability. In both cases, Tanner tried to resolve the discrepancy by raising the contextual priority for one of the contrasting concept projections. From the KiP perspective, Tanner might activate a non-relevant knowledge element to raise the contextual priority for some concept projection; however, this does not mean that a decline in his knowledge system has happened. This is part of a continuous learning process in which he reduces the contextuality until he reaches a level at which he can apply any relevant concept projection about any facet of sampling variability productively in any context. What he needs is an exceedingly wider span of different data contexts using the same concept projection before he develops expertise (coordination class) for the concept of sampling variability. Thinking in different contexts involves constructing new concept projections with a continuous alignment process that leads to deactivating any inaccurate or non-relevant knowledge elements (Wagner, 2006). Tanner's difficulties reported in this study, however, have not been shown by any previous studies especially using the KiP framework from the cognitive science. That is, this study went beyond De Vetten et al. (2018) which investigated PSTs' understanding for the idea of making conclusions that go beyond the data at hand and understanding the representativeness of the sample for the population.

This study also shows that even detailed reasoning about sampling variability in one data context may not be sufficient to say that a PST has a complete understanding of this notion. This finding is consistent with the prior work on reasoning in mathematics, physics, and statistics done from the Knowledge in Pieces perspective (e.g., Levin, 2018; Izsák & Jacobson, 2017; Wagner, 2006, 2010) in which contextuality of reasoning processes is one of the central assumptions. However, the goal of this study is not to show that contextuality affects PSTs' reasoning about sampling variability, but to use some contextuality episodes to reveal the difficulties that PSTs face while they reason about this concept and understand what is behind these difficulties. The unique aspect contributed by this study is the use of the techniques of Knowledge in Pieces from the learning sciences to describe at a moment-by-moment level the cognitive dynamics of the transitions that occurred in the PST's knowledge system. This has provided detailed specifications of PSTs' difficulties in reasoning about sampling variability by describing some of the knowledge elements that have supported and constrained their abilities to reason about the concept of sampling variability. A strength of applying the Knowledge in Pieces to this data analysis was the facility that enabled an account for PSTs' seemingly contradictory or irrelevant knowledge elements related to a given facet of the concept of sampling variability. One goal of the data analysis in this study was to show that the coordination classes model (diSessa & Sherin, 1998; diSessa & Wagner, 2005; diSessa et al., 2016) can be used effectively to investigate the concept of sampling variability. By doing so, more research from the learning sciences arena might investigate learning this concept in depth.

While some of the research on sampling variability has been focused on either the misconceptions learners have or their lack of appropriate knowledge (Reading & Shaughnessy, 2004; Meletiou–Mavrotheris & Lee, 2002), this study promotes the idea that learners difficulties might not necessarily arise from lack of knowledge, but from the activation of less-productive knowledge elements over others. This is in line with many studies that re-called "misconceptions" as the activation of knowledge elements that are not effective for the given context, but that are not necessarily "wrong" (Elby & Hammer, 2010; Hammer, 2000; Smith et al., 1993). As the episodes presented in this analysis show, some knowledge elements—even if they are not usually activated—can be helpful when brought to Tanner's attention in one particular data context because they might be combined and coordinated; therefore, they might likely be used productively by Tanner in an increasingly wider span of data contexts.

As mentioned before, PSTs' difficulties in reasoning about the concept of sampling variability have been documented across some studies. The question becomes what can we do to meaningfully support PSTs' reasoning through instruction. For this, it is essential to know about the ways in which PSTs' reasoning about this concept is contextually grounded. We need to talk to them, gather data, and look at the way they reason. From that, we build appropriate instructional plans and materials. When PSTs experience disequilibrium in their knowledge systems about the concept of sampling variability because of the way that they construe some data context and the features they attend to, the disequilibrium does not always result in a more normative organization of their knowledge systems as seen in the above analysis. Opportunities should be provided to engage PSTs in multiple tasks and data contexts, and careful follow-up questions need to be asked by the teacher to know what exactly happened in their knowledge systems as a result of the experienced disequilibrium. The case of Tanner clearly exhibited that although he wrestled with disequilibrium for some time, this did not help him determine which was the statistically accurate reasoning. While experiencing disequilibrium may be an important instructional moment, it is far from clear yet how to most effectively leverage such an experience or whether this is always the most desirable intervention. Although the interviews occurred over a short span of time (approximately 60 minutes), I was able to elicit in a targeted way the phenomenon of contextuality in PSTs' reasoning about the concept of sampling variability. Foremost, the scope of this study is to capture some contextuality episodes in PSTs' reasoning about sampling variability, then to analyze these episodes and offer conjectures for what is behind them. Therefore, this study cannot make any claims as to the frequency that the observed difficulties would occur within the overall PSTs population; larger-scale statistical studies would be needed to shed light on this question.

References

Ben-Zvi, D., Gil, E., & Apel, N. (2007). What is hidden beyond the data? Helping young students to reason and argue about some wider universe. In *Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5)*. University of Warwick, UK.

Ben-Zvi, D., Makar, K., Garfield, J., & Eds. (2018). *International Handbook of Research in Statistics Education*. Springer International Handbooks of Education. https://doi.org/10.1007/978-3-319-66195-7

Brown, N. J. S., Danish, J. A., Levin, M., & diSessa, A. A. (2015). Competence reconceived: The shared enterprise of knowledge analysis and interaction analysis. In A. A. diSessa, M. Levin, & N. J. S. Brown (Eds.), *Knowledge and interaction: A synthetic agenda for the learning sciences*. Routledge.

Browning, C., Goss, J., & Smith, D. (2014). Statistical knowledge for teaching: Elementary preservice teachers. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the 9th International Conference on Teaching Statistics, Flagstaff, AZ, USA*. International Statistical Institute.

Burgess, T. A. (2011). Teacher knowledge of and for statistical investigations. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics–challenges for teaching and teacher education (pp. 259–270)*. Springer.

Canada, D., & Ciancetta, M. (2007). *Elementary preservice teachers' informal conceptions of distribution*. Proceedings of the 29th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, NV.

Canada, D., & Makar, K. (2006). Preservice teachers' informal descriptions of variation. In *2006 AERA Annual Meeting: Research in the Public Interest, San Francisco, California, 7-11 April, 2006*. Online: American Educational Research Association.

Canada, D. L. (2008). Conceptions of distribution held by middle school students and preservice teachers. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE study: Teaching statistics in school mathematics. Challenges for teaching and teacher education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference*. ICMI and IASE.

Cobb, G., & Moore, D. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801–823.

Cobb, P. (2007). Putting philosophy to work. Coping with multiple theoretical perspectives. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 3–38). Information Age Publishing.

Confrey, J. (1986). A critique of teacher effectiveness research in mathematics education. *Journal for Research in Mathematics Education*, *17*, 347–360.

De Vetten, A., Schoonenboom, J., Keijzer, R., & Van Oers, B. (2018). Pre-service primary school teachers' knowledge of informal statistical inference. *Journal of Mathematics Teacher Education*, *22*(6), 639–661. https://doi.org/10.1007/s10857-018-9403-9

delMas, R. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 79–96). Kluwer Academic.

diSessa, A. (1988). Knowledge in pieces. In G. Forman & P. Putall (Eds.), *Constructivism in the computer age* (pp. 49–70). Lawrence Erlbaum Associates.

diSessa, A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*, 105–225.

diSessa, A. A. (1996). What do “just plain folk” know about physics? In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching, and schooling* (pp. 709–730). Blackwell Publishers.

diSessa, A. A. (2018). A friendly introduction to “Knowledge in Pieces”: Modeling types of knowledge and their roles in learning. In G. Kaiser, H. Forgasz, M. Graven, A. Kuzniak, E. Simmt, & B. Xu (Eds.), *Invited Lectures from the 13th International Congress on Mathematical Education, ICME-13 Monographs* (pp. 66–84). Springer Open. https://doi.org/10.1007/978-3-319-72170-5_5

diSessa, A. A., Sherin, B., & Levin, M. (2016). Knowledge analysis: An introduction. In A. diSessa, M. Levin, & N. Brown (Eds.), *Knowledge and interaction: A synthetic agenda for the learning sciences* (pp. 30–71). Routledge.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change. *International Journal of Science Education*, *20*, 1155–1191.

diSessa, A. A., & Wagner, J. F. (2005). What coordination has to say about transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 121–154). Information Age Publishing.

Elby, A., & Hammer, D. (2010). Epistemological resources and framing: A cognitive framework for helping teachers interpret and respond to their students’ epistemologies. In L. Bendixen & F. Feucht (Eds.), *Personal epistemology in the classroom: Theory, research, and implications for practice* (pp. 409–434). University Press.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre K–12 curriculum framework*. American Statistical Association.

Gil, E., & Ben–Zvi, D. (2010). Emergence of reasoning about sampling among young students in the context of informal inferential reasoning. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence–based society (Proceedings of the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia, July 11–16)*. International Statistical Institute.

Groth, R. E., & Bergner, J. A. (2006). Preservice elementary teachers’ conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning*, *8*(1), 37–63.

Hahn, C. (2014). Linking academic knowledge and work experience in using statistics, a design experiment for business school students. *Educational Studies in Mathematics*, *86*(2), 239–251.

Hammer, D. (2000). Student resources for learning introductory physics. *American Journal of Physics, Physics Education Research Supplement*, *68*(S1), S52–S59.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards Mathematics*. Washington D.C.

Izsák, A., & Jacobson, E. (2017). Preservice teachers’ reasoning about relationships that are and are not proportional: A knowledge–in–pieces account. *Journal for Research in Mathematics Education*, *48*(3), 300–339.

Jacobbe, T., & Carvalho, C. (2011). Teachers’ understanding of averages. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Joint ICMI/IASE study: Teaching statistics in school mathematics. Challenges for teaching and teacher education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference* (pp. 199–209). Springer.

Jones, S. R. (2013). Understanding the integral: Students’ symbolic forms. *The Journal of Mathematical Behavior*, *32*(2), 122–141.

Kapon, S., & diSessa, A. A. (2012). Reasoning through instructional analogies. *Cognition and Instruction*, *30*(3), 261–310.

Konold, C., & Miller, C. D. (2015). *TinkerPlots: Dynamic data exploration [Computer software, Version 2.3]*. Learn Troop.

Levin, M. (2018). Conceptual and procedural knowledge during strategy construction: A complex knowledge systems perspective. *Cognition and Instruction*, *36*, 247–278.

Makar, K., Bakker, A., & Ben–Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*(1–2), 152–173.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105.

Meletiou–Mavrotheris, M., & Lee, C. (2002). Teaching students the stochastic nature of statistical concepts in an introductory statistics course. *Statistics Education Research Journal*, *1*(2), 22–37.

Meletiou–Mavrotheris, M., & Paparistodemou, E. (2015). Developing students’ reasoning about samples and sampling in the context of informal inferences. *Educational Studies in Mathematics*, *88*(3), 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.), *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9)*. Voorburg, the Netherlands.

Noll, J. A. (2011). Graduate teaching assistants' statistical content knowledge of sampling. *Statistics Education Research Journal*, *10*(2), 48–74.

Pfannkuch, M. (2008). Building sampling concepts for statistical inference: A case study. *Proceedings of ICME–11*, Monterrey, Mexico, July 2008

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

Pfannkuch, M., Arnold, P., & Wild, C. J. (2015). What I see is not quite the way it really is: Students' emergent reasoning about sampling variability. *Educational Studies in Mathematics*, *88*(3), 343–360.

Reading, C., & Shaughnessy, J. M. (2004). Reasoning about variation. In D. Ben–Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 201–226). Kluwer Academic Publishers.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). NCTM.

Smith, J., diSessa, A., & Roschelle, J. (1993/1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*, 115–163.

Smith, J. P. (1995). Competent reasoning with rational numbers. *Cognition and Instruction*, *13*, 3–50.

Triola, M. F. (2018). *Elementary statistics* (13th ed.). Pearson.

van Dijke–Droogers, M., Drijvers, P., & Bakker, A. (2020). Repeated sampling with a black box to make informal statistical inference accessible. *Mathematical Thinking and Learning*, *22*(2), 116–138.

von Glasersfeld, E. (1991). Abstraction, re–presentation, and reflection. In L. P. Steffe (Ed.), *Epistemological foundations of mathematical experience* (pp. 45–67). Springer.

Vosniadou, S. (2013). Conceptual change in learning and instruction: The framework theory approach. In S. Vosniadou (Ed.), *International Handbook of Research on Conceptual Change* (pp. 11–30). Routledge.

Wagner, J. F. (2006). Transfer in pieces. *Cognition and Instruction*, *24*(1), 1–71.

Wagner, J. F. (2010). A transfer–in–pieces consideration of the perception of structure in the transfer of learning. *Journal of the Learning Sciences*, *19*(4), 443–479.

Watson, J. M., & Callingham, R. (2013). Likelihood and sample size: The understandings of students and their teachers. *The Journal of Mathematical Behavior*, *32*(3), 660–672.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.