Educational Studies in Mathematics (2023) 114:17–34
https://doi.org/10.1007/s10649-023-10234-z

A method for assessing students' interpretations of contextualized data

Randall E. Groth[1] · Yoojin Choi[1]

Accepted: 13 April 2023 / Published online: 13 May 2023

Abstract

Learning to interpret data in context is an important educational outcome. To assess students' attainment of this outcome, it is necessary to examine the interplay between their contextual and statistical reasoning. We describe a research method designed to do so. The method draws upon Toulmin's (1958, 2003) model of argumentation for the first stage of qualitative data analysis and the Structure of the Observed Learning Outcome (SOLO) (Biggs & Collis, 1991) model for the second stage. Toulmin analyses help identify the justifications and expressions of uncertainty students provide in their interpretive arguments. Subsequent analyses based on the multi-modal conceptualization of SOLO help characterize the quality of student arguments relative to one another. Existing literature and an empirical example are drawn upon to explain how the Toulmin and SOLO models can be used in tandem to analyze students' interpretations of contextualized data. We also explain how pairing Toulmin and SOLO can address theoretical and practical limitations that arise when using just one of the two models on its own.

Keywords Qualitative research · SOLO taxonomy · Toulmin model · Statistics · Argumentation · Context knowledge

In applied statistics, data and context are inextricably linked, as reflected in Cobb and Moore's (1997) maxim that "data are not just numbers, they are numbers with a context" (p. 801). Wild and Pfannkuch (1999) noted that during statistical investigations, there needs to be "continual shuttling backwards and forwards between thinking in the context sphere and the statistical sphere" (p. 228). Knowledge of the contexts in which data were generated and statistical knowledge are both needed to construct interpretations. Plausible interpretations of statistics often rely upon knowledge of the context in which the data were generated, just as knowledge of a given context can be gained by examining statistics derived from it. Konold and Higgins (2003) characterized this complementary knowledge generation process as a "give-and-take conversation" in which context knowledge enhances statistical data analysis and data analysis enhances knowledge of context.

✉ Randall E. Groth
regroth@salisbury.edu

[1] Department of Secondary and Physical Education, Salisbury University, 1101 Camden Ave., Salisbury, MD 21801, USA

Given the importance of coordinating statistical and contextual knowledge, teaching students to do so has become an increasingly prominent instructional goal. Ben-Zvi and Aridor-Berger (2016) noted the need for teachers to help students traverse between the data and context worlds and understand how to integrate the two. Similarly, Bargagliotti et al. (2020) called for students' interpretations of statistical investigations to "integrate the context and objectives of the investigation to draw conclusions from data and to support these conclusions using statistical evidence" (p. 103). A growing number of teaching strategies with the potential to help students coordinate statistical and contextual knowledge have been developed, implemented, and analyzed (Makar & Ben-Zvi, 2011; Pfannkuch et al., 2018). As such teaching strategies emerge, it is useful to have methods that assess the extent to which they attain their goals for student learning. Such methods provide information that can be used to continuously improve teaching strategies and to compare student learning outcomes associated with different curricula.

1 Purpose and structure of the article

The purpose of this article is to present a method that can be used to assess students' interpretations of contextualized data. To introduce the method, we begin by describing salient cognitive dynamics of interpreting contextualized data. We explain how Toulmin's (1958, 2003) argumentation model can be used during the first stage of analysis of students' interpretations. We then explain how the Structure of the Observed Learning Outcome (SOLO) model (Biggs & Collis, 1991) can provide the basis for a second stage of analysis in which students' statistical interpretations are compared against one another. To conclude, limitations and delimitations of our proposed two-stage method are discussed.

2 Cognitive dynamics of interpreting contextualized data

Students' interpretations of data provide windows into their reasoning about integrating statistical and contextual knowledge. Interpretations are essential components of statistical investigations; they present claims about questions related to groups or populations from which data are drawn (Bargagliotti et al., 2020). Because interpretations are responses to statistical questions, they must be stochastic rather than deterministic in nature; limitations to generalizability and degrees of uncertainty due to sample-to-sample variability, missing data, bias, and other statistical and contextual factors need to be acknowledged (Bargagliotti et al., 2020). Providing justifications while acknowledging limitations is a complex cognitive task. As Ben-Zvi et al. (2012) put it, doing so requires students to navigate a "middle ground between knowing everything and knowing nothing" (p. 923); providing justifications without limitations may render an interpretation that is deterministic and overly certain in nature, and focusing solely on limitations and uncertainty can lead to a relativistic conclusion that no firm knowledge can be derived from a contextualized data set. Hence, methods for assessing students' statistical interpretations need to consider students' abilities to balance justifications of their claims about data against appropriate expressions of uncertainty and qualifications. In this section, we consider research on students' justifications as well as ways in which they can acknowledge limitations in qualifying their interpretations.

2.1 Students' justifications of statistical interpretations: evidential and abductive reasoning

Evidential reasoning and abductive reasoning (Gil & Ben-Zvi, 2011) are both needed to interpret contextualized data. They can be described in the following manner:

> *Evidential reasoning* is the process of arriving at the inference that justifies why the data should be regarded as appropriate evidence in support of the inference and weighing the strength of the evidence. *Abductive reasoning* is the process of providing a contextual or theoretical support for the data being as they are (Gil & Ben-Zvi, 2011, p. 92).

Evidential and abductive reasoning both contribute to plausible interpretations of contextualized data when deployed effectively in conjunction with one another. For example, Gil and Ben-Zvi described the case of two students who produced graphs which appeared to indicate that sixth-graders had greater long-jump distances than seventh-graders. This surprised the two students and contradicted their context-related expectation that the older children would be able to jump further. Subsequently, they re-analyzed the original data, and they found that there were far fewer girls than boys in the sixth-grade group, which helped explain the counter-intuitive statistical result. In this situation, evidential reasoning (using data, graphs, and statistics) and abductive reasoning (using context knowledge related to the long-jump) were both needed to help students create a plausible interpretation of the data.

Instances of students' evidential and abductive reasoning complementing one another can be found in other studies as well. Langrall et al. (2011) found that students tended to use context-related reasoning to explain patterns they observed in data. For example, some students in their research provided the story behind a set of World Cup data by using contextual knowledge of penalties that were assessed during games and players whose performances contributed heavily to the observed team data. Similarly, Shaughnessy and Pfannkuch (2002) described how students' background knowledge of the *Old Faithful* geyser helped them explain patterns in eruption data and choose optimal representations to justify predictions about when eruptions would occur. Makar and Rubin (2009) found that contextual and statistical knowledge students gained by gathering and analyzing handspan data from their classmates helped them make plausible predictions about handspan data that might be gathered from similar classes. In such cases, evidential and abductive reasoning work in tandem to produce richer interpretations than would be formed by using just one of the two types of reasoning alone.

Although abductive reasoning can help students form and justify interpretations of contextualized data, research has also shown that using context knowledge productively can be challenging. Students' context knowledge can, at times, introduce considerations that are irrelevant to the statistical question at hand. For example, Langrall et al. (2011) found that students' context knowledge about the World Cup sometimes led to extended discussions that did not move the investigation of a group's statistical question forward. Contextual considerations can also divert learners' attention from statistical concepts a teacher may wish to develop. Pfannkuch (2011) described a case in which students' inventive stories about the origins of unusual heights in a data set made it difficult for the teacher to refocus a lesson on the statistical ideas of sample, population, sample size, and sampling variability. Additionally, students' inaccurate beliefs about context can sometimes prevent them from forming viable interpretations of data. Masnick et al. (2007), for example, found that some children's initial incorrect beliefs about factors influencing pendulum motion remained in place even after they had gathered and analyzed data contradictory to the initial beliefs.

Collectively, such studies indicate that helping students coordinate evidential and abductive reasoning to form plausible statistical interpretations is a non-trivial, yet vital, task.

2.2 Students' acknowledgment of limitations of statistical interpretations: expressing uncertainty

Plausible interpretations of contextualized data generally cannot be stated in absolute terms. Uncertainty due to statistical and contextual factors makes it necessary to qualify most claims. As students begin to learn to interpret contextualized data, they can use qualitative language to express uncertainty about the claims they make during statistical investigations, though doing so can be complicated. Ben-Zvi et al. (2012) illustrated the complexity of helping students find a middle ground between deterministic and relativistic statistical claims. Initially, students in their research tended to make absolute statements about data they were asked to explore. For example, when asked to compare data on communication preferences of boys and girls, students made claims such as "only the girls like to talk by cell phone" and "boys don't like cell phones at all" (p. 918). Later, when given a small set of data, students went to the opposite extreme of believing that nothing could be claimed from the data. Eventually, students did start to use the language of uncertainty more effectively, as when they proposed investigating the question, "What do girls usually like?" (p. 918; introducing "usually" as an appropriate qualifier). However, deterministic and relativistic interpretations still emerged at times. Similarly, Henriques and Oliveira (2016) noticed that some students expressed too much confidence in how characteristics of a sample would generalize to a larger population, while others used appropriate qualitative language, such as "probably," "maybe," or "tend to be," to qualify their statistical interpretations.

The process of learning to qualify statistical interpretations appropriately has both qualitative and quantitative elements. As students use qualitative language to qualify claims, it is important for them to recognize that qualitative words can be arranged along a continuum according to the levels of certainty they convey (Groth et al., 2020). For example, "probably" expresses a greater degree of certainty than "unlikely" and "never." After learning to select words that correspond to levels of certainty appropriate for a given situation, students can eventually learn to quantify the degree of certainty as well (Bargagliotti et al., 2020). As with selecting appropriate qualitative words to qualify interpretations, using quantitative qualifiers can also be challenging. Even professionals who use statistics in their field at times have trouble doing so. For example, LeMire (2010) argued that much of the controversy surrounding null hypothesis statistical testing (NHST) is due to lack of acknowledgement of its limitations rather than because of NHST itself. Type I and type II errors are among the concepts that can and should be used to qualify formal inferential claims produced through NHST, but they are often not deployed effectively. School curricula must be purposeful in how they develop students' abilities to qualify statistical interpretations both qualitatively and quantitatively.

3 Toulmin's model and students' interpretations of statistical data

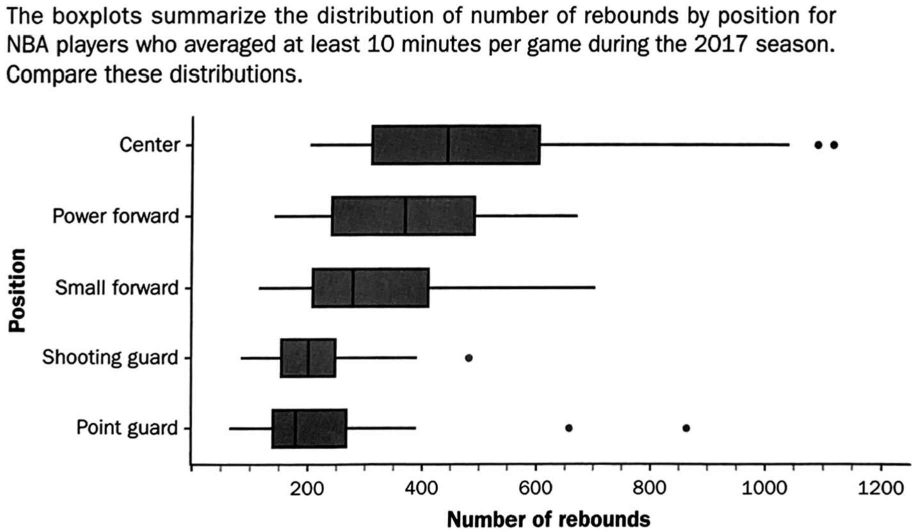
Given the statistics education research we have discussed to this point, theoretical frameworks that undergird analyses of students' interpretations of data need to account for the justifications students offer and the ways they express limitations and uncertainties.

Toulmin's (1958, 2003) model of argumentation includes these elements, so we draw upon it as a theoretical basis for the first stage of the data analysis method presented in this article. Using Toulmin's model to assess students' statistical reasoning is not unique to our work. The model has previously been used to analyze arguments related to both formal inference (e.g., LeMire, 2010) and informal inference (e.g., Gil & Ben-Zvi, 2011). We do, however, propose a new direction by showing how Toulmin analyses of students' statistical arguments are compatible with the Structure of the Observed Learning Outcome (SOLO) model, which is frequently used in statistics education research. Next, we explain the components of the Toulmin model and provide an empirical example of its application before showing how the analysis can be extended and enhanced with SOLO.

Toulmin's (1958, 2003) model contains six primary components: data, warrant, claim, backing, qualifiers, and rebuttals. *Data* are facts or evidence that provide the foundation for a *claim*. Data alone, however, provide just one component of a convincing argument. One also needs a *warrant* which links the data and claim. Additionally, strong arguments tend to have *backing* to support the warrant. In mathematics education research, warrant and backing are often referred to collectively as *justification* (e.g., Chazan et al., 2012; González & Eli, 2017). We adopt this convention for pragmatic reasons, as it eliminates the need to parse warrant from backing, which is a well-documented difficulty in applying the Toulmin model (Warren, 2010). Toulmin's model, which is not limited to using formal logic to map argument structures, is flexible enough to allow for justifications to be evidential or abductive in nature. Toulmin's model is silent, however, about evaluating the quality of the justifications offered, leaving that task to domain-specific standards capable of doing so (Nussbaum, 2011).

Toulmin's (1958, 2003) model also includes *qualifiers* and *rebuttals* as components of arguments. An argument/statistical interpretation becomes more plausible and trustworthy as one acknowledges its potential limitations. Qualifiers such as "possibly," "probably," and "usually" help specify the extent of an interpretive claim's applicability. Past research on students' use of the language of uncertainty during statistical investigations provides several examples of qualifiers, such as "probably," "maybe," "tend to be," "usually," and "majority" (Ben-Zvi et al., 2012; Henriques & Oliviera, 2016). Langrall et al. (2011) found that students with context expertise for a given task were more likely to qualify statistical claims than those who did not have such expertise. Using qualifiers in everyday arguments helps one avoid over-stating a claim; in statistics, qualifiers help express degrees of uncertainty when one communicates an interpretation to others. *Rebuttals* are counter-arguments that acknowledge exceptions to claims. An example of a rebuttal in a statistics classroom can be seen in the dialogue that Ben-Zvi et al. (2012) described as students discussed the extent to which findings from classroom data about students' free-time activities might apply to larger populations. Some of the students acknowledged that other classrooms might yield substantially different results and hence limit the strength of claims that could be made. From a Toulmin perspective, such acknowledgements are resonant with rebuttals, which identify situations for which one's overarching interpretive claim about data may not hold.

To further illustrate the application of Toulmin's model to the analysis of students' statistical interpretations, we use responses gathered to the task shown in Fig. 1, which is from a text on using statistical reasoning in sports contexts (Tabor & Franklin, 2019). We asked 11 physical education majors to write responses to it as part of a pilot study for a larger project. They ranged from 19 to 22 years of age, and according to university records, 10 were male and 1 was female. We use this task as an example because it allowed the respondents to bring context expertise (Langrall et al., 2011; Mooney, 2002) related to their



The boxplots summarize the distribution of number of rebounds by position for NBA players who averaged at least 10 minutes per game during the 2017 season. Compare these distributions.

Fig. 1 Basketball task (Tabor & Franklin, 2019, p. 255). From *Statistical Reasoning in Sports* 2e by Josh Tabor and Christine Franklin. Copyright W.H. Freeman 2019. All rights reserved. Used by permission of the publisher Macmillan Learning

academic major to bear both to justify and to qualify their interpretations. We had them complete the task in one of their physical education classes rather than in a mathematics or statistics class, conjecturing that this would further encourage the use of context expertise. Respondents had not formally studied argumentation structures. Although the task incorporates boxplots, which are among the most difficult representations to interpret (Bakker et al., 2004; Edwards et al., 2017), boxplots were included in the respondents' pre-university curricula (Common Core State Standards Initiative, 2010), and four of them encountered boxplots again during an introductory university-level statistics course. The participants' backgrounds and the structure of the Fig. 1 task allowed us to observe the expression of both evidentiary and abductive justifications as well as limitations to arguments.

Next, we discuss one of the more complex responses we received to the Fig. 1 task to illustrate how multiple Toulmin components can be contained in a given student interpretation. Participant S7's response, with our Toulmin-based annotations included, was:

> Point guards average the least rebounds but have a bigger range than shooting guards. {Some point guards have more rebounds than power and small forwards}. **Centers average the most** [because that is *usually* what they're there for]. Small forwards average less than power forwards but have a bigger range.

In the above excerpt, we considered the main overarching claim to be about how positions ranked in relation to one another in terms of number of rebounds (the phrase "centers average the most" is rendered in bold print above to summarize this claim). Evidential (single underline above) and abductive justifications (text enclosed in square brackets) were both included. In the evidential realm, averages were mentioned in support of the ranking given, though S7 was not specific about the type of average (it is not clear if S7 considered the median to be a type of average, as in some statistics texts, or misinterpreted the vertical lines in the boxes to indicate means rather than medians). S7 used abductive reasoning to explain patterns in the data, as reflected in the observation that it is usually the center's responsibility to rebound (text enclosed in square brackets above). S7 cited some unusual data points in order to acknowledge exceptions to the overarching claim, observing that some point guards have more rebounds than power and small forwards. Acknowledging exceptions to a claim is indicative of a rebuttal (text enclosed in curly brackets above). S7 introduced the qualifier "usually" (italic print above) to acknowledge that rebounding may, at some times, not be the center's responsibility. S7 also offered another claim about variation in referring to ranges; however, here, we focus on the larger overarching claim about how positions tend to rank regarding rebounds. Toulmin analyses of arguments are often summarized by using diagrams like the one shown in Fig. 2, which summarizes our analysis of S7's response. The data and overarching claim appear in the first row of the diagram, and justifications, rebuttals, and qualifiers appear beneath it.

Diagramming one of the most complex responses first provides a useful baseline for analysis of other responses. Each subsequent response can be concisely summarized by editing the initial diagram (in this case, Fig. 2), removing and adding elements as needed. Examples of responses we received that yielded different diagram configurations are shown in Table 1. In our small data set, only one other response (from S8) contained a claim, evidential justification, abductive justification, qualifiers, and rebuttals, and hence had a very similar Toulmin diagram. S8's abductive justification differed from S7's in that S8 cited the heights of centers, rather than their positional responsibilities, to explain their greater numbers of rebounds. One response (from S9) offered a claim with evidential and abductive justification but no qualifiers or rebuttals. Three responses (S1, S3, and S6) consisted of a claim, abductive justification, and qualifiers. Another response (S4) had a claim, evidential justification, and qualifiers. S5 responded with contextual considerations and a qualifier, but no clear comparative claim. Some responses (S2, S10, S11) were simply claims without explicit justification or qualification. Each response can be represented by adding elements to Fig. 2, removing them, or editing them. Collaboratively editing a relatively complex diagram to represent other such responses from a small initial data set provides an opportunity to build a shared understanding of how Toulmin elements should be operationalized for a given task. Research teams with large sets of responses to analyze can then use their shared understanding to inform independent analyses and then compare their resultant Toulmin diagrams afterward.

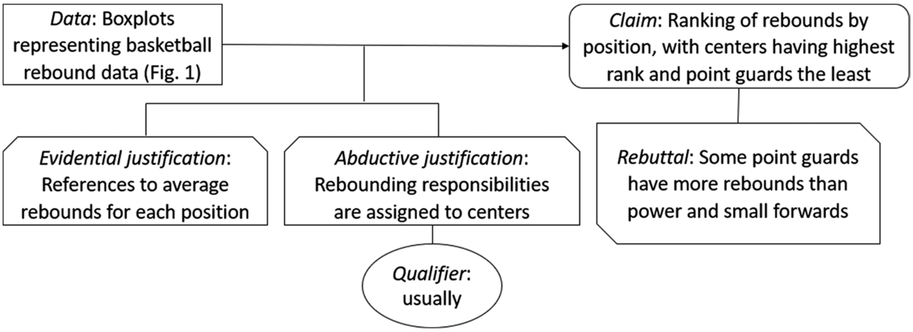

Data: Boxplots representing basketball rebound data (Fig. 1)
Claim: Ranking of rebounds by position, with centers having highest rank and point guards the least
Evidential justification: References to average rebounds for each position
Abductive justification: Rebounding responsibilities are assigned to centers
Rebuttal: Some point guards have more rebounds than power and small forwards
Qualifier: usually

Fig. 2 Toulmin diagram summary of S7's response

Table 1 Examples of coded arguments consisting of different combinations of Toulmin model elements

Participant	Argument elements	Coded response
S8	**claim**, evidential justification, [abductive justification], *qualifiers*, {rebuttals}	The **center gets the most** [because he is the tallest on the court]…but the small forward *might* {get more in a couple games}. Shooting and point guard average the same [because they are the shortest ones on the court].
S9	**claim**, evidential justification, [abductive justification]	**Center – highest number of rebounds**, two players with above average rebounds. Point guard – lowest number of rebounds, with two players above average as well. [Centers are bigger and closer to the rim…]
S1	**claim**, [abductive justification], *qualifiers*	**Centers obtain the most rebounds** per game, followed by power forward. Centers and power forwards *typically* get the most rebounds [because they are the biggest on the court].
S4	**claim**, evidential justification, *qualifiers*	**Centers** *typically* **had the highest amount of rebounds** overall with the highest average and stats. Power forwards had the second highest average. Small forwards had the third highest average. Shooting guards and point guards had the lowest rebounding averages with around 180–200 rebounds.
S5	[abductive contextual considerations], *qualifiers*	My guess is that because of [positioning on the court] that players have better opportunities at those spots *at times.*
S11	**claim**	**The point guard and shooting guard have a lot less rebounds than the other positions.**

4 SOLO analyses of statistical interpretation argument structures

Although Toulmin analyses provide starting points for characterizing the complexity of arguments, they do not provide the final word on the quality of responses in relation to one another. The Toulmin model was designed to apply to multiple fields of endeavor, so domain-specific standards must also be brought to bear when assessing the relative quality of arguments (Nussbaum, 2011). In the domain of statistics education, the Structure of the Observed Learning Outcome (SOLO) model (Biggs & Collis, 1991) is frequently used to assess students' responses to tasks (Langrall et al., 2017). SOLO has helped researchers assess individuals' responses to tasks involving distributions (Reading & Reid, 2006), measures of center (Groth & Bergner, 2006), variation (Watson et al., 2022b), statistical investigations (Pfannkuch, 2005), and many other statistical ideas (Jones et al., 2004). SOLO-based studies help teachers and researchers anticipate the levels of response they may encounter as students complete academic tasks (Watson, 2006). Having a common structure such as SOLO across studies also facilitates the process of synthesizing research to produce increasingly comprehensive frameworks to describe statistical thinking (Jones et al., 2000; Mooney, 2002). Next, we discuss how Toulmin and SOLO analyses can complement one another in assessing students' interpretations of contextualized data.

SOLO is a Neo-Piagetian model that consists of five modes of development that are similar (but not precisely equivalent) to Piaget's (1983) stages. The modes Biggs and Collis (1991) posited in describing SOLO were: sensorimotor, ikonic, concrete symbolic, formal, and post-formal. Statistics education researchers using the SOLO model have predominantly encountered student responses indicative of the ikonic and concrete symbolic modes (Groth et al., 2021; Jones et al., 2004). The former mode is characterized primarily by intuitive knowledge and mental imagery and the latter by the use of written language and symbols (Biggs & Collis, 1991; Pegg, 2014). Watson et al. (2022b) described characteristics of the ikonic and concrete symbolic modes in the context of investigating students' understanding of variation. In one task, students were asked to interpret data they collected and plotted about plant growth. Concrete symbolic interpretations drew upon data and statistics from one or more specific variables that were studied, such as height of the plant, treatment, or number of days elapsed. This sort of support for claims is resonant with evidential reasoning. Ikonic mode responses tended to pertain to details about the context of the activity, such as recollections and imagery of the actions they had to take in measuring and collecting plant growth data. Using such contextual details to explain patterns in data is resonant with abductive reasoning.

Pairing abductive ikonic mode reasoning with concrete-symbolic evidential reasoning can result in exceptionally rich responses to tasks. The SOLO model emphasizes the potential value of such multi-modal reasoning. Biggs and Collis (1991) postulated that early modes of development are not necessarily replaced by later ones; rather, ideas characteristic of earlier modes sometimes play supportive roles in solving problems related to later modes of development. Biggs and Collis gave an example of how a scientist's work modeling the structure of organic ring compounds was supported by a personal ikonic mental image of six snakes chasing each other. Similarly, Groth et al. (2021) explained how ikonic mode thinking tendencies pertaining to problem context, such as visualizing the positions of objects in a container, forming images of random generators, and thinking about past experiences playing games of chance, can support students' learning. Although ikonic mode thinking and contextual knowledge can at times be distracting and lead students astray when interpreting data (Jones et al., 2000; Langrall et al., 2011; Pfannkuch, 2011), it is important to help students learn to coordinate the two effectively. Multi-modal reasoning that effectively coordinates evidential and abductive reasoning is necessary in order to do the shuttling back and forth between data and context (Wild & Pfannkuch, 1999) required of statisticians.

In the small set of responses we gathered for the Fig. 1 task, S7 and S8 coordinated evidential and abductive reasoning in productive ways, even though their responses did not rise to the level one would expect from professional statisticians. Abductive reasoning about contextual aspects such as players' heights and position responsibilities helped explain the data and statistics represented in the boxplots. In these cases, contextual experiences playing and watching the game (accessible through the ikonic mode) supported deeper analysis of the data and accompanying representations and statistics (accessible through the concrete-symbolic mode). Essentially, knowledge gained from basketball experiences supported the process of "reading behind the data" (Shaughnessy, 2007) to identify patterns in data and judge whether they were typical of what one would expect to see in larger data sets and those from similar contexts (e.g., claims about centers generally having the most rebounds, point guards the least). In Fig. 3, we depict this sort of multi-modal functioning as a goal for students' learning to interpret statistical data. Not all responses in our small data set coordinated abductive and evidential reasoning in these ways. Some focused solely on contextual factors to justify their responses (e.g., player height, position responsibilities), and others relied solely on statistical information (e.g., position of

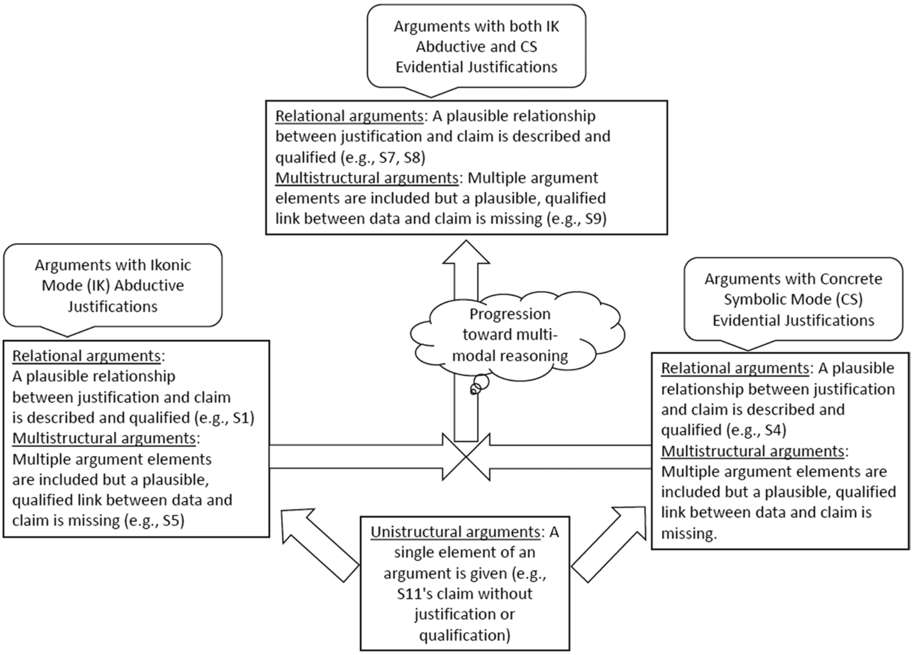

Arguments with both IK Abductive and CS Evidential Justifications
Relational arguments: A plausible relationship between justification and claim is described and qualified (e.g., S7, S8)
Multistructural arguments: Multiple argument elements are included but a plausible, qualified link between data and claim is missing (e.g., S9)
Progression toward multi-modal reasoning
Arguments with Ikonic Mode (IK) Abductive Justifications
Relational arguments: A plausible relationship between justification and claim is described and qualified (e.g., S1)
Multistructural arguments: Multiple argument elements are included but a plausible, qualified link between data and claim is missing (e.g., S5)
Arguments with Concrete Symbolic Mode (CS) Evidential Justifications
Relational arguments: A plausible relationship between justification and claim is described and qualified (e.g., S4)
Multistructural arguments: Multiple argument elements are included but a plausible, qualified link between data and claim is missing.
Unistructural arguments: A single element of an argument is given (e.g., S11's claim without justification or qualification)

Fig. 3 Hypothetical progression toward using multi-modal reasoning to construct interpretations of contextualized data. Descriptions of potential unistructural, multistructural, and relational levels are adapted from the Watson et al. (2022a) investigation of multi-modal functioning in statistics and the Pezaro et al. (2014) synthesis of the Toulmin and SOLO models. The sample responses that are referenced in connection with different levels are shown in Table 1 and Fig. 2

boxplots relative to one another, averages). Some offered claims without any type of justification. Figure 3 depicts responses that reflect just one mode of reasoning as being earlier steps on a progression toward richer, multi-modal reasoning. The categories in Fig. 3 provide an initial frame of reference for comparing the types of arguments students produce when interpreting statistical data in response to a contextualized task, while emphasizing that developing students' ability to employ multi-modal reasoning when interpreting data is an important instructional goal.

The SOLO model also suggests the possibility of organizing responses into finer-grained levels of sophistication within each of the categories depicted as rectangles in Fig. 3. In the SOLO model, modes contain cycles consisting of three levels: unistructural, multistructural, and relational (Biggs & Collis, 1991). Many SOLO-based statistics education research studies use unistructural-multistructural-relational (UMR) cycles to qualitatively analyze students' responses to statistical tasks and further compare the responses to one another in terms of levels of sophistication (Langrall et al., 2017). At the unistructural level, responses attend to a single aspect of relevance for a given task. Multistructural responses contain multiple aspects of relevance, but the aspects are not tied together with a unifying coherent explanation. A unifying explanation is apparent in relational responses. Watson et al. (2022a) detected UMR cycles in student responses that reflected the ikonic mode, the concrete-symbolic mode, and a combination of the two. Such potential to have UMR cycles within each mode of reasoning, as well as having UMR cycles in multi-modal reasoning, is acknowledged in the rectangles in Fig. 3. Although not explicitly depicted in Fig. 3, it should also be noted that there can be multiple UMR cycles within a given mode (Pegg, 2014; Watson et al., 1995).

Having students' interpretations of data represented as Toulmin diagrams is a useful way to begin to discern UMR cycles within a set of qualitative student responses. Pezaro et al. (2014) demonstrated how the SOLO and Toulmin models could be used in tandem for this purpose. Their work was done in the context of studying prospective teachers' scientific arguments. The Toulmin model was used to map argument structures, and SOLO was used to characterize the sophistication of the structures relative to one another. The number of Toulmin model elements present in each argument and their degrees of interconnectedness were used in order to place responses at appropriate SOLO levels. So, for example, unistructural responses might contain just one Toulmin model element, leading to claims that were not well-supported. Multistructural and relational responses might contain multiple Toulmin elements, but they would be adequately interconnected to form a coherent argument only at the relational level. In sum, Pezaro et al. (2014) assigned SOLO levels according to the quantity of relevant Toulmin elements present in an argument and the quality of element integration. Integrating Toulmin elements such as qualifiers, rebuttals, and justifications can contribute to the coherence of an argument.

Doing a detailed, fine-grained UMR cycle analysis is most plausible when working with a large set of student response data (e.g., Watson & Moritz, 1998). In smaller sets of data, responses representative of some important UMR levels may not be present. Given the small preliminary data set of responses we gathered for the Fig. 1 task, we did not conduct a detailed UMR level analysis of the responses we received. However, some of the student interpretations for the task did bear resemblance to unistructural, multistructural, and relational responses, as noted in Fig. 3. Some responses contained just one element of the Toulmin model: a claim about how the distributions compared to one another, and no additional details to justify or qualify the claim were offered. S11, for example, wrote, "The point guard and shooting guard have a lot less rebounds than the other positions." Such responses resemble the unistructural level of SOLO, in which just one element of relevance to the given task is offered (in S11's case, a claim). S5's response offered a relevant abductive consideration and qualifier, but it did not include a clear claim. S5 wrote, "My guess is that because of positioning on the court that players have better opportunities at those spots at times." The presence of multiple Toulmin elements, but absence of a plausible unifying link between data and claim, is resonant with multistructural reasoning in SOLO. In contrast, S7 and S8 (see Table 1) offered claims that were accompanied by relevant justifications and qualifiers. The justifications and qualifiers enhanced the coherence of the arguments and helped tie their elements together. The presence of multiple relevant elements that are woven together to produce a coherent narrative is akin to the relational level in the SOLO model. The extent to which these sample responses match UMR levels in SOLO could come into sharper relief as researchers compare them against additional responses from large qualitative student response data sets. As this is done, researchers might also find multiple UMR cycles in multi-modal reasoning. In more advanced UMR cycles, perhaps respondents would be more precise than S7 and S8 in specifying the statistics and contextual details that support their interpretations.

5 Delimitations and limitations related to pairing Toulmin and SOLO to analyze students' interpretations of statistical data

As researchers consider pairing the Toulmin and SOLO models to analyze students' interpretations of data in the ways we have described, it is important to consider delimitations and limitations of doing so. There are delimitations pertaining to the scope and focus of studies that could potentially be carried out with the method, and there are limitations pertaining to potential weaknesses of studies carried out with this method. Regarding delimitations, we discuss the scope of student tasks to which the method applies, its emphasis on learning outcomes, and its cognitive focus. Regarding limitations, we discuss scholarly critiques of the SOLO and Toulmin models that pertain to their falsifiability, emphasis on abstract knowledge, and ambiguities they introduce during qualitative data analysis.

5.1 Delimitations: student tasks, learning outcomes, and cognitive focus

Pairing the Toulmin and SOLO models narrows the scope of research that can be done in comparison to using just one of the two on its own. SOLO is most suitable for analyzing students' responses to open-ended tasks that allow for multiple approaches reflective of a wide range of reasoning. The Toulmin model is designed specifically to analyze argument structures. Hence, the qualitative data analysis strategy we have described has somewhat narrow applicability to open-ended tasks that call for arguments. Nonetheless, understanding students' thinking in relation to these types of tasks is a high priority for statistics education researchers. There have been consistent calls for statistics instruction to include opportunities for data-informed decision-making rather than just mere computation of statistics and production of graphs (Bargagliotti et al., 2020; Shaughnessy, 2007). The need for intelligent data-informed decision-making was brought into sharp relief by the COVID-19 pandemic, as statistical arguments became a common part of public discourse (da Silva et al., 2021; Kollosche & Meyerhöfer, 2021; Rubel et al., 2021). By carefully analyzing the arguments individuals make to themselves and others about the implications of data and statistics, we can more clearly define priorities for statistics curricula. Our pairing of Toulmin and SOLO is well-suited to this specific type of analysis.

Another delimitation is that the Toulmin and SOLO pairing is mainly applicable to assessing learning outcomes rather than processes. As the final letter in the SOLO acronym suggests, the model assesses *outcomes* of learning, and the Toulmin model is often used to analyze complete structures of arguments offered. In using SOLO, Chick (1998) observed, "the complexity exhibited in the observed outcome can differ from that of the cognitive processes used to achieve that outcome" (p. 20). Chick suggested that it is useful to examine the intermediate steps students take to produce a response, rather than just the finished response itself, to understand more fully their cognitive processes. In our empirical illustration, for instance, it would have been interesting to know how respondents developed certain elements of their contextual knowledge of basketball. For example, there are at least two ways to conclude that centers tend to have the most rebounds. One might develop an intuition about centers and rebounding after playing or watching games; alternatively, one might make conclusions about centers' rebounding by examining data. The former is more indicative of ikonic mode functioning, and the latter of concrete symbolic. Given the possibility of multi-modal functioning, a combination of the two modes of thinking might also be used. We obtained some information to indicate how respondents justified their arguments, but supplementary methods that incorporate observations, interviews, and other ways to monitor student learning as it develops would be needed to help investigate the development of cognitive processes and argument structures.

Our empirical example is like many other studies that use SOLO or Toulmin to focus on the outcomes of individuals' cognitive processes, so the method we propose is not highly compatible with studies based on theoretical frameworks that de-emphasize or reject the notion of individual cognition. However, it should also be noted that using SOLO and Toulmin together does not completely prohibit the analysis of arguments that individuals collaboratively construct in a social context. Although our empirical example did not illustrate the analysis and comparison of collaboratively constructed arguments, others have used Toulmin or SOLO for these purposes. For example, Watson et al. (1995) assessed the SOLO levels of statistical analyses conducted by small groups of students. Group-constructed responses were assigned SOLO levels, and learning outcomes attained by different groups were compared to one another. Similarly, mathematics education researchers have used and adapted the Toulmin model to evaluate group-constructed arguments in several instructional contexts (e.g., Chazan et al., 2012; González & Eli, 2017). Knipping and Reid (2015) used Toulmin's model to construct global portraits of argumentation that permitted comparisons between classrooms. In statistics education research, it would be useful to compare statistical interpretations constructed in classrooms that incorporate various developing pedagogies such as growing samples (Ben-Zvi et al., 2012), provocative tasks (Madden, 2011), extended contextualized investigations (Watson et al., 2022b), and different modeling approaches (Pfannkuch et al., 2018). Although most studies using SOLO and Toulmin focus on individuals' responses (as exemplified by our empirical illustration), it is nonetheless plausible to seek ways to combine the two models to study group-constructed interpretive arguments as well. Doing so may be of interest to researchers investigating the contributions that students with statistical expertise and those with context expertise make to collaborative interpretations of statistical data (e.g., Langrall et al., 2011).

5.2 Limitations: falsifiability, analytic ambiguities, and abstract knowledge emphasis

Some limitations of the qualitative data analysis method we propose can be traced to its theoretical underpinnings. Shaughnessy (2007) summarized some of the scholarly critiques of SOLO, stating, “One of the criticisms leveled against the SOLO model is that it is not falsifiable, so the validity of any conclusions reached via a SOLO approach cannot be easily challenged” (p. 1001). Shaughnessy also claimed that “the SOLO model is based on the assumption that development can be represented in hierarchical structures” and questioned the warrant for such an assumption. This line of critique implies that statistics education researchers have not put sufficient effort into scrutinizing the validity of the SOLO model, or that the model itself does not lend itself to the possibility of change based on empirical observation. Pairing Toulmin and SOLO helps address such concerns by allowing researchers to engage in theory triangulation (Denzin, 2009; Schoenfeld, 2008). Conducting Toulmin analyses before SOLO analyses allows for the possibility of finding interpretive arguments that do not fit well into modes or levels suggested by SOLO. This, in turn, puts researchers in position to suggest changes to the SOLO model, or, in more extreme cases, reject SOLO altogether in favor of a more adequate model to assess the relative quality of responses. Of course, theoretical blind spots shared by both theories will remain, but the number of blind spots can be reduced by this sort of theoretical cross-validation.

Another limitation of the method we propose relates to the scholarly critique that analyzing qualitative data with SOLO can be ambiguous and lead to coding disagreements among researchers. Chan et al. (2002) provided examples to illustrate this concern. In some cases, Chan et al. found that raters categorized a given response at the highest level in a SOLO-based rubric, and others categorized it at the lowest level. During UMR analyses, raters at times disagreed about what constituted a listing of ideas (i.e., multistructural) as opposed to a generalization or integration of ideas (i.e., relational). To resolve such ambiguities, researchers need mechanisms for identifying the number of relevant elements in a response and assessing the coherence with which the elements are woven together. The Toulmin model provides a means for researchers to work toward a common understanding of these aspects of SOLO (Pezaro et al., 2014). Relevant elements can be defined as the Toulmin components built into the response (claims, justifications, qualifiers, etc.). A response might be considered unistructural if it contains just one of these elements, and possibly multistructural if it contains more than one. Relational responses might be conceived of as those in which a claim is both offered and supported with relevant justification and qualification. Teams of researchers conducting SOLO analyses can work toward common understanding of how to characterize student responses by initially analyzing responses collaboratively, as we did for the responses we received to the Fig. 1 task. Initial collaborative work can provide a basis for later independent coding, if desired. Of course, ambiguities will still arise under such a scenario, but they can be reduced by using Toulmin model components as a common vocabulary to construct clearer shared definitions of essential SOLO aspects.

SOLO has also received scholarly critique for valuing abstract knowledge to such an extent that concrete knowledge gained from experiences in everyday contexts is inordinately de-emphasized (Kahn, 2015). The use of the phrase “extended abstract” to describe higher-level student responses in foundational SOLO work (Biggs & Collis, 1982) would seem to fuel this critique. Attending to this critique is particularly important in statistics education, as both context and mathematical abstractions are needed for coherent interpretations of data (Wild & Pfannkuch, 1999). As Cobb and Moore (1997) noted, “In data analysis, context provides meaning” (p. 803); this contrasts with purely deductive mathematical thinking, where “the context is part of the irrelevant detail that must be boiled off over the flame of abstraction to reveal the previously hidden crystal of pure structure” (p. 803). Efforts to understand students’ statistical thinking in context (e.g., Langrall et al., 2011; Pfannkuch, 2011) require frameworks that account for the roles played by both context knowledge and abstract mathematical knowledge.

The multi-modal conceptualization of SOLO (e.g., Biggs & Collis, 1991; Watson et al., 1995; Watson et al., 2022a), developed after initial foundational SOLO work (Biggs & Collis, 1982), helps address critiques related to over-emphasis of abstract knowledge. According to the multi-modal conceptualization, the ikonic mode is not replaced by the concrete-symbolic mode; rather, both are acquired by learners so that ikonic mode thinking remains available after the concrete-symbolic mode is acquired. So, abductive reasoning developed in the ikonic mode (e.g., from playing or watching a sport) can be used in tandem with evidential reasoning characteristic of the concrete-symbolic mode. Interpretations of contextualized data are usually enhanced by drawing upon both types of reasoning. It is necessary, however, to carefully assess the quality of reasoning brought to bear by each mode. Abductive reasoning can at times lead students astray by introducing elements such as myths, superstitions, and imaginative stories to interpretations of data (Groth et al., 2021). Evidential reasoning can include misconceptions about statistical concepts, or, as in our empirical example, vague statements about the specific statistics used to arrive at an interpretation. Although the Toulmin/SOLO combination we described is compatible with examining how evidential and abductive reasoning work in tandem, it cannot, on its own, detect weaknesses in arguments that stem from statistical or contextual considerations. Hence, when using the method we propose, it is necessary for researchers also to draw upon statistical and contextual knowledge related to student tasks when analyzing responses to them. Sometimes this will require having those with context expertise on research teams work alongside those with statistical expertise, as in our case, where one author was from the field of statistics education and the other from physical education.

6 Conclusion

Theory-based methods are essential to the infrastructure for doing research. We have described how the methodological infrastructure for statistics education research can be enhanced by using the Toulmin model in conjunction with the multi-modal conceptualization of SOLO. This two-stage qualitative data analysis method provides research infrastructure for addressing the high-priority task of understanding students' reasoning when merging knowledge of statistics and context to interpret contextualized data. The two-stage method also helps address theoretical and practical limitations that emerge when using either the Toulmin or SOLO model on its own. We hope the qualitative data analysis method we have described continues to evolve and develop as researchers examine its applicability to assessing students' interpretations of data from various contexts.

Declarations

Competing interests The authors declare no competing interests.

References

Bakker, A., Biehler, R., & Konold, C. (2004). *Should young students learn about box plots?* IASE Roundtable, Lund, Sweden. http://www.statlit.org/PDF/2004BakkerIASE.pdf. Accessed 6 Apr 2023.

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K-12 guidelines for assessment and instruction in statistics education II*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf. Accessed 6 Apr 2023.

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

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

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press. https://doi.org/10.1016/C2013-0-10375-3

Biggs, J. B., & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behavior. In H. A. H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 57–66). Lawrence Erlbaum Associates.

Chan, C. C., Tsui, M. S., Chan, M. Y. C., & Hong, J. H. (2002). Applying the Structure of the Observed Learning Outcomes taxonomy on students' learning outcomes: An empirical study. *Assessment and Evaluation in Higher Education, 27*(6), 511–527. https://doi.org/10.1080/0260293022000020282

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

Chick, H. (1998). Cognition in the formal modes: Research mathematics and the SOLO taxonomy. *Mathematics Education Research Journal, 10*(2), 4–26. https://doi.org/10.1007/BF03217340

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801–823. https://doi.org/10.2307/2975286

Common Core State Standards Initiative (2010). *Common core state standards for mathematics*. http://www.corestandards.org/. Accessed 6 Apr 2023.

da Silva, A. S., Barbosa, M. T. S., de Souza Velasque, L., da Silveira Barroso Alves, D., & Nascimento Magalhães, M. (2021). The COVID-19 epidemic in Brazil: How statistics education may contribute to unravel the reality behind the charts. *Educational Studies in Mathematics, 108*(1–2), 269–289. https://doi.org/10.1007/s10649-021-10112-6

Denzin, N. K. (2009). *The research act: A theoretical introduction to sociological methods*. Routledge. https://doi.org/10.4324/9781315134543

Edwards, T. G., Özgün-Koca, A., & Barr, J. (2017). Interpretations of boxplots: Helping middle school students to think outside the box. *Journal of Statistics Education, 25*(1), 21–28. https://doi.org/10.1080/10691898.2017.1288556

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

González, G., & Eli, J. A. (2017). Prospective and in-service teachers' perspectives about launching a problem. *Journal of Mathematics Teacher Education, 20*(2), 159–201. https://doi.org/10.1007/s10857-015-9303-1

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. https://doi.org/10.1207/s15327833mtl0801_3

Groth, R. E., Bergner, J. A., & Austin, J. W. (2020). Dimensions of learning probability vocabulary. *Journal for Research in Mathematics Education, 51*(1), 75–104. https://doi.org/10.5951/jresematheduc.2019.0008

Groth, R. E., Austin, J. W., Naumann, M., & Rickards, M. (2021). Toward a theoretical structure to characterize early probabilistic thinking. *Mathematics Education Research Journal, 33*(2), 241–261. https://doi.org/10.1007/s13394-019-00287-w

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

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning, 2*(4), 269–307. https://doi.org/10.1207/S15327833MTL0204_3

Jones, G. A., Langrall, C. W., Mooney, E. S., & Thornton, C. A. (2004). Models of development in statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 97–117). Kluwer. https://doi.org/10.1007/1-4020-2278-6_5

Kahn, P. (2015). Critical perspectives on methodology in pedagogic research. *Teaching in Higher Education, 20*(4), 442–454. https://doi.org/10.1080/13562517.2015.1023286

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

Kollosche, D., & Meyerhöfer, W. (2021). COVID-19, mathematics education, and the evaluation of expert knowledge. *Educational Studies in Mathematics, 108*(1–2), 401–417. https://doi.org/10.1007/s10649-021-10097-2

Konold, C., & Higgins, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193–215). National Council of Teachers of Mathematics.

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

Langrall, C. W., 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.

LeMire, S. D. (2010). An argument framework for the application of null hypothesis statistical testing in support of research. *Journal of Statistics Education*, *18*(2). https://doi.org/10.1080/10691898.2010.11889492

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

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105. https://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1)_Makar_Rubin.pdf. Accessed 6 Apr 2023.

Makar, K., & Ben-Zvi, D. (2011). The role of context in developing reasoning about informal statistical inference. *Mathematical Thinking and Learning, 13*(1&2), 1–4. https://doi.org/10.1080/10986065.2011.538291

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

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning, 4*(1), 23–64. https://doi.org/10.1207/S15327833MTL0401_2

Nussbaum, E. M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist, 46*(2), 84–106. https://doi.org/10.1080/00461520.2011.558816

Pegg, J. (2014). Structure of the Observed Learning Outcome (SOLO) model. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 570–572). Springer. https://doi.org/10.1007/978-94-007-4978-8_182

Pezaro, C., Wright, T., & Gillies, R. (2014). Pre-service primary teachers' argumentation in socioscientific issues. *Proceedings of the Frontiers in Mathematics and Science Education Research Conference* (pp. 58–69). Famagusta, North Cyprus. https://www.scimath.net/download/pre-service-primary-teachers-argumentation-in-socioscientific-issues-9627.pdf. Accessed 6 Apr 2023.

Pfannkuch, M. (2005). Characterizing year 11 students' evaluation of a statistical process. *Statistics Education Research Journal, 4*(2), 5–26. https://doi.org/10.52041/serj.v4i2.512

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

Pfannkuch, M., Ben-Zvi, D., & Budgett, S. (2018). Innovations in statistical modeling to connect data, chance, and context. *ZDM-Mathematics Education, 50*(7), 1113–1123. https://doi.org/10.1007/s11858-018-0989-2

Piaget, J. (1983). Piaget's Theory. In P. Mussen (Ed.), *Handbook of child psychology* (pp. 103–128). John Wiley & Sons.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal, 5*(2), 46–68. https://doi.org/10.52041/serj.v5i2.500

Rubel, L. H., Nicol, C., & Chronaki, A. (2021). A critical mathematics perspective on reading data visualizations: Reimagining through reformatting, reframing, and renarrating. *Educational Studies in Mathematics, 108*(1–2), 249–268. https://doi.org/10.1007/s10649-021-10087-4

Schoenfeld, A. H. (Ed.). (2008). A study of teaching: Multiple lenses, multiple views. *Journal for Research in Mathematics Education monograph series (Vol. 14)*. Reston, VA: National Council of Teachers of Mathematics.

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* (Vol. 2, pp. 957–1009). Information Age Publishing and National Council of Teachers of Mathematics.

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

Tabor, J., & Franklin, C. (2019). *Statistical reasoning in sports* (2nd edn). Bedford, Freeman, & Worth.

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

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

Warren, J. E. (2010). Taming the warrant in Toulmin's model of argument. *The English Journal*, *99*(6), 41–46. https://www.jstor.org/stable/20787665. Accessed 6 Apr 2023.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Lawrence Erlbaum Associates.

Watson, J. M., & Moritz, J. B. (1998). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145–168. https://doi.org/10.1023/A:1003594832397

Watson, J. M., Collis, K. F., Callingham, R. A., & Moritz, J. B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation, 1*(3), 247–275. https://doi.org/10.1080/1380361950010303

Watson, J., Fitzallen, N., & Wright, S., & Kelly, B. (2022a). Characterizing student experience of variation within a STEM context: Improving catapults. *Statistics Education Research Journal*, *21*(1), Article 9. https://doi.org/10.52041/serj.v21i1.7

Watson, J. M., Wright, S., Fitzallen, N., & Kelly, B. (2022b). Consolidating understanding of variation as part of STEM: Experimenting with plant growth. *Mathematics Education Research Journal.* Advance online publication. https://doi.org/10.1007/s13394-022-00421-1

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

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