CHARACTERIZING STUDENT UNDERSTANDING OF VARIATION WITHIN A STEM CONTEXT: IMPROVING CATAPULTS

JANE WATSON
University of Tasmania
jane.watson@utas.edu.au

NOLEINE FITZALLEN
University of Tasmania
noleine.fitzallen@utas.edu.au

SUZIE WRIGHT
University of Tasmania

BEN KELLY
University of Tasmania
ben.kelly@utas.edu.au

ABSTRACT

STEM learning experiences at the school level provide both opportunities and challenges for exploring students' understanding of statistical concepts. This report focuses on data handling and informal inference embedded in a STEM context, that is, of testing, adjusting, and retesting catapults. In particular, the learning goal was for Grade 4 students (aged 9–10 years) to build on their developing understanding of variation while learning about the science topic of force as demonstrated by two configurations of catapults causing ping pong balls to be launched different distances. This report focuses on the students' experiences of variation that were associated with the activity from a structural perspective during implementation. The analysis, employing various aspects of the Structure of Observed Learning Outcomes, points to the potential contribution of Multimodal functioning in identifying and characterizing understanding of variation in a new context. The activity took place with 50 students in two classes, with data collected from student workbooks and graphs created in TinkerPlots. Results suggest that meaningful engagement with context can provide support for developing understanding of the concept of variation.

Keywords: *Statistics education research; Variation; SOLO model; Practice of statistics; STEM education; TinkerPlots*

1. INTRODUCTION

1.1. BACKGROUND

The strong interest of governments in Science, Technology, Engineering, and Mathematics (STEM) fields, both individually and collectively, as critical for national economic growth and competitive advantage (e.g., Engler, 2012; Office of the Chief Scientist, 2013), has put increasing pressure on education systems to provide graduates prepared to work in these fields. At the same time, the STEM-related fields of Big Data and Data Science have emerged (e.g., Finzer, 2013; François & Monteiro, 2018; Ridgway et al., 2018). Together these movements impact on the compulsory years of schooling in preparing students to be able to take advantage of STEM careers in the future, careers that are likely to be influenced by the power of data and statistics to drive change and innovation (Watson et al., 2020a). The issue is how to make the most of STEM learning opportunities for young students in terms of addressing the need to develop their statistical knowledge and data skills across the breadth of STEM disciplines. The challenge lies in finding learning activities that are embedded within STEM learning contexts that require motivating, meaningful data collection opportunities, with each complementing the other.

Statistics Education Research Journal, 21(1), Article 9. https://doi.org/10.52041/serj.v21i1.7

The degree of integration possible in education across the STEM disciplines depends on many factors, including the curriculum expectations at the various grade levels in the four related fields. In Australia, these fields are Science, Digital Technology, Design Technology, and Mathematics (Australian Curriculum, Assessment, and Reporting Authority [ACARA], 2019). It should be noted that the Engineering discipline of STEM is subsumed in the Design Technology curriculum, whereas Engineering is included in the *Next Generation Science Standards* (National Research Council, 2013) in the United States. Fundamental, however, to any STEM activity involving data collection is the practice of statistics (Moore & McCabe, 1989; Watson et al., 2018). The practice of statistics encompasses the concepts covered by the summary of statistical problem-solving in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report* (Franklin et al., 2007, pp. 11–12):

- Formulate question/s, anticipating variability:
    - Clarify the problem at hand
    - Formulate one (or more) questions that can be answered with data
- Collect data, designing for variability:
    - Design a plan to collect appropriate data
    - Employ the plan to collect the data
- Analyze data, accounting for variability:
    - Select appropriate graphical and numerical methods
    - Use these methods to analyze the data
- Interpret results, allowing for variability:
    - Interpret the analysis
    - Relate the interpretation to the original question

The more recent *GAISE II Report* (Bargagliotti et al., 2020, pp. 13–15) uses the same framework emphasizing variability and indicating the possible cyclic and non-linear nature of the problem-solving process with the four fundamental steps (reproduced in Figure 1).

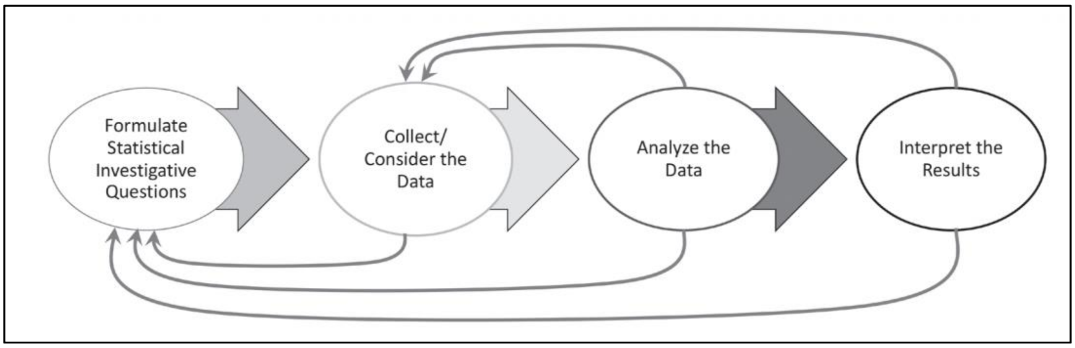


Figure 1. GAISE II Statistical problem-solving (Bargagliotti et al., 2020, p. 13)

As seen in the *GAISE* description, the fundamental concept underlying the practice of statistics is variation, without which there would be no need for statistics (e.g., Moore, 1990). Three other essential concepts, besides variation, also underpin the practice of statistics. They are distribution, the basis for the analyze step in *GAISE*; expectation, the notion that leads to asking questions and collecting data; and informal inference, the approach to decision-making that acknowledges uncertainty (Watson et al., 2018). In terms of variation, distribution is the lens through which variation can be viewed (Wild, 2006), expectation is the attempt to harness variation for a purpose (Watson, 2005), and informal inference acknowledges the uncertainty associated with results obtained (Makar & Rubin, 2009). All four concepts are embedded within statistical inquiries set within STEM contexts. For children, as for adults working in applied STEM fields, the goal is to explain the variation experienced to tell a story or answer a question. Little research, however, has taken place specifically linking variation and STEM contexts in relation to student outcomes.

Appreciation of students' development of understanding of variation received very little attention in the early days of statistics education research. Green (1993) and Shaughnessy (1997) began asking questions about what students knew about variation and how understanding originated. They also asked for positive evidence of what students were able to do, rather than just identifying their deficiencies. Individual interviews and surveys of students began to be employed across grade levels and longitudinally, with Shaughnessy et al. (1999) interviewing 324 students at three sites across two countries in Grades 4 to 12 about sampling from a container with 100 lollies, of which 50 were red, 30 green, and 20 yellow. The study showed growth across grade levels on the center (expectation) criteria of the task but no clear improvement on the spread (variation) criteria. Similarly, Noll and Shaughnessy (2012) provided an extensive account of the development of reasoning about variation in six middle and high school classes. Their hierarchy of understanding began with additive thinking, progressing to using one attribute (e.g., range, most, gap) of a distribution or the population, then to distributional thinking combining two or more single attributes to describe potential expectation within variation.

Surveys and interviews were used across various ages of students over the next decade to explore students' development of understanding of variation in relation to the other fundamental concepts associated with statistical investigations. These studies, for example, Watson and Callingham (2003), Watson et al. (2007), and Watson (2009), based their analyses on the Structure of Observed Learning Outcomes (SOLO) model (Biggs & Collis, 1982, 1991). SOLO was used to develop hierarchical models of student performance, generally in terms of the relationships between variation and expectation. Watson et al. (2003), for example, developed a survey that was administered to 746 students across Grades 3 to 9. The initial coding was based on SOLO, which was followed by a Rasch analysis that suggested four levels of development of reasoning were evident: (a) Prerequisites for variation (e.g., simple graph reading, intuitive chance reasoning), (b) Partial recognition of variation (e.g., ideas in context, focusing on single aspects only), (c) Applications of variation (e.g., consolidating ideas in context but inconsistent on salient features), and (d) Critical aspects of variation (e.g., complex justification, critical reasoning). The survey items were also used to measure change related to teaching interventions; for example, Watson and Kelly (2004) reported improvement across grades from pre- to post-tests and longitudinal tests after lessons on chance and data, for items testing understanding of variation with respect to spinners. Other reports used the surveys to measure change in student outcomes where variation occurred in chance, graphing, and sampling situations in Grade 3 (Watson & Kelly, 2002a), Grade 5 (Watson & Kelly, 2002b), and Grades 7 and 9 (Watson & Kelly, 2002c) in association with teaching interventions. These studies provided the background for measuring progress on the development of understanding of variation as a foundation for observing application of understanding while carrying out hands-on investigations for the current study.

Petrosino et al. (2003) reported on a teaching intervention that comprised a series of eight lessons with a Grade 4 class based on variation to introduce distributions and their properties, including typical and spread, using three contexts of measuring: the height of a flagpole, the length of a pencil, and the heights reached by rockets with two different types of nose cones. Different sources of variation were introduced in each context across the lessons. The flagpole was measured by two different instruments to yield different degrees of variation. Measuring the pencil demonstrated the decrease in variation of measurements in a context with more control of the process. Measuring the heights when firing two types of rockets introduced a second source of variation, with the expectation of one rocket reaching a greater height than the other. The class discussion demonstrated the students' progress in quantifying variation to consider difference in spread for the two methods of measuring the height of the flagpole. The context was then extended for rockets to consider the variation in expected height for the two methods of launching the rockets.

Reading (2004) posed a context of planning an outdoor festival to Grade 7, 9, and 11 students, for which they had to choose a month of the year based upon data for either rainfall or temperature for the month. Variation within one of the annual data sets was the focus of the analysis by the students. Reading analyzed the reasoning displayed in responses during the activity rather than in a post assessment, finding that qualitative understanding of the context often influenced predictions made. Based on the analysis, she hypothesized two hierarchical cycles of response within the Concrete Symbolic mode of the SOLO taxonomy (Biggs & Collis, 1982), one based on qualitative descriptions of the context and the other on quantitative descriptions employing the numerical information available.

More recently, Watson et al. (2020b) initiated a project, including the study reported here, with an activity about making "licorice" by hand and by machine. The focus of the activity was on the variation in the data when using the two different methods of production. The expectation was the licorice sticks would be the *same* size and mass for both methods. The variation explored in the activity was evident in the the data produced by the two methods: The machine-made licorice pieces were quite consistent in mass, whereas the mass of the hand-made licorice pieces varied greatly. Again, student responses were coded based on the SOLO model. This activity had a similar aim to that of Petrosino et al. (2003) in comparing two methods of measuring the height of a flagpole. The contexts helped students begin to appreciate the nuances of variation without having to deal simultaneously with difference in expectation.

As can be seen in these examples, previous research has embedded the study of students' understanding of variation in various contexts related to measurement data as well as discrete numerical data, much of it descriptive of student experience in the classroom. Little research, however, has taken place with students working through extended structured learning experiences specifically within the larger statistical problem-solving arena of the *GAISE Reports* (Bargagliotti et al., 2020; Franklin et al., 2007). The overall aim of the current study was to contribute to knowledge about students' experiences of variation from exposure to a STEM context within which data were collected and analyzed. This research expands the work done by Petrosino et al. (2003) that involved launching two different types of rockets, with the expectation of variation in the outcomes.

1.2. THEORETICAL PERSPECTIVE

Following some of the research noted in the previous section, the theoretical model chosen for the analysis of student responses in this study was the neo-Piagetian SOLO model of Biggs and Collis (1982, 1989, 1991). The SOLO model proposes five modes of development: Sensorimotor (objects and space in a perceived environment, physical interaction, curiosity), Ikonic (qualitative reasoning, intuition, imagery), Concrete Symbolic (quantitative reasoning, everyday thinking, concrete logical systems), Formal (deductive thinking with abstract concepts and rules, formalisation of knowledge), and Post Formal (new abstractions, basic research). Of interest in this study are the Ikonic mode, which arises at around 18 months, and the Concrete Symbolic mode, starting from around 6 years of age. As the name implies, the Ikonic (IK) mode draws on what is perceptually evident (1991, p. 72), but also perhaps on myth-making and intuitive thought. The Concrete Symbolic (CS) mode moves to symbolization of written language and the mathematical symbol system and their application in real world contexts. Within each of the modes, generally there are potentially three levels of reasoning observed in responses, depending on how many aspects or elements of a task are employed in the responses: Unistructural (U), where a single element or idea contributes to reasoning; Multistructural (M), where reasoning includes two or more elements presented in a serial fashion; and Relational (R), where reasoning includes links or relationships among multiple elements presented. Although much of the research on school-level mathematics has taken place in the CS mode, there have often been examples of young students providing responses considered to be in the IK mode. In reporting on student responses related to chance, Watson et al. (1997) and Watson and Moritz (1998) found many examples of IK responses, often related to superstitious beliefs or favourite colours; therefore, they included a general IK-level with the analyses of the predominantly CS responses. Based on interviews with students across grade levels from early childhood to grade 10, in which students created pictographs from a story with data and then interpreted the story and made predictions (Watson, 2007), Watson and Moritz (2001) observed many IK responses, often in tandem with CS responses and related to variation.

A natural extension of studying multiple SOLO cycles and modes is the consideration of Multimodal functioning (Collis & Romberg, 1991), which involves the interaction of IK and CS modes. Although the modes are hierarchical, Biggs and Collis (1991) argued that the CS mode does not replace the IK mode. Instead, there can be facilitation of CS learning by referring to IK experiences. Biggs and Collis (1989) suggested each mode potentially "remains as an option for learning throughout life" with Multimodal learning extending through all higher modes (p. 160). Earlier studies involving fractions (Watson et al., 1993) and higher order statistical thinking by primary school students (Watson et al., 1995) illustrated students' capacity to exhibit Multimodal functioning, particularly including the IK mode. The qualitative and observational nature of IK comments may establish or reinforce a related CS response, or evidence a superstition that may be detrimental to CS thinking. Watson and Collis (1994) provide multiple examples of the positive influence of IK contributions for students being interviewed on tasks in chance and data contexts. For example, when viewing data representations or creating them from concrete materials, some students made judgements based on overall physical size before proceeding to count as a way of justifying their qualitative visual observations.

Recent work by Groth et al. (2021), describing young students' probabilistic thinking, extended the characterization of the IK mode by specifically suggesting two parallel U-M-R cycles in the mode and providing labels for the two potential types of thinking involved. One is termed normative *incompatible*, characterized by myths, superstitions, and subjective or deterministic beliefs. The other, termed normative *compatible*, includes personal experiences, such as imagery and intuition, which although not CS in nature, are in the appropriate context for the expectations of the CS mode and could be supporting the normative thinking of that mode. Groth et al. proposed similar characteristics for the parallel U-M-R levels of the IK mode as for the other modes, in that students either use normative-incompatible thinking or normative compatible thinking in attending to one aspect of the task at hand (U), attending to multiple aspects of the task (M), or weaving a consistent narrative with multiple aspects of the task (R).

Combining the Multimodal work of Collis and Romberg (1991), Watson et al. (1995), and Watson and Collis (1994), with the Groth et al. (2021) nomenclature and use of U-M-R cycles in the IK mode, produces multiple problem-solving pathways as shown in Figure 2. Moving vertically there are three potential U-M-R pathways to a problem solution. The horizontal arrows in the figure also indicate the IK support that may be provided to a CS response to a task, in a Multimodal fashion. In this study there is potential for the extended SOLO model to be used to characterize the progression of thinking about variation with respect to a STEM activity using catapults; in particular, there is the potential of IK contextual thinking to encourage and support CS numerical responses in student work.

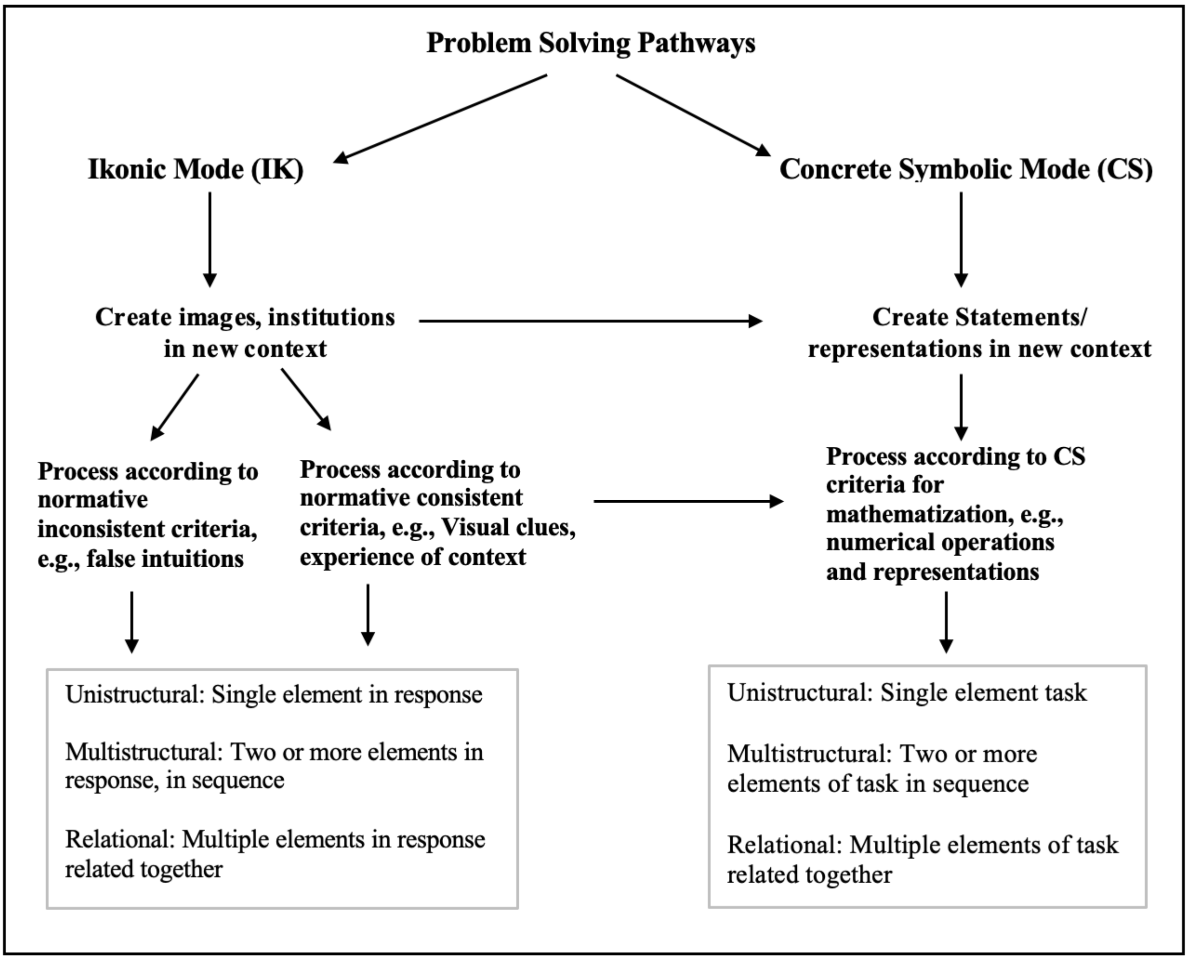


Figure 2. Potential problem-solving pathways

1.3. RESEARCH QUESTION

Of particular interest in this study is the identification of students' conceptualisations of variation embedded within a STEM context, which is represented by the data they collect, and how the variation in the data is viewed through distributions, to make inferences in relation to the questions asked within the activity. The research question reflects this goal:

> How can students' understanding of variation during a STEM inquiry be characterized in terms of the problem-solving pathways (Figure 2)?

2. RESEARCH METHOD

2.1. CONTEXT OF THE STUDY AND RESEARCH APPROACH

The research reported in this paper arises from the second year of a longitudinal project that explored young students' development of understanding of statistical concepts set within STEM learning activities, as they progressed from Grade 3 to Grade 6 (aged 8–12 years). The activity reported here was the fourth, carried out in the second half of the academic year when the students were in Grade 4 (aged 9–10 years). The research was considered interpretive in nature and employed predominantly qualitative research strategies (Creswell, 2013) to capture students' thinking as they experienced the activity. Such an approach was necessary to ensure the data revealed the different ways in which the students made sense of and used data during the statistical inquiry that was based on the practice of statistics (Watson et al., 2018).

A design-based research approach (Cobb et al., 2003) was employed using the results from the earlier activities in the longitudinal study to inform the implementation of activities that followed. The aim was to build competence in carrying out the practice of statistics and to focus on different fundamental concepts and steps of the practice in successive activities, which were first introduced to this student cohort in Grade 3 (Chick et al., 2018; Watson et al., 2020b). Taking the design-based research approach (Cobb et al., 2003), the STEM activity reported in this paper extended the students' experience with variation to include a STEM context within which the expectation of performance was *different* for two iterations of data collection. To serve this purpose, the topic of Catapults was chosen for the student activity.

2.2. THE STUDENT ACTIVITY: CATAPULTS

The Catapult activity involved launching ping pong balls from pre-fabricated catapults, making adjustments to the catapults, and relaunching ping pong balls to determine if the adjustments improved the performance of the catapults. Improving the performance of the catapults was designed to be a motivating context within which to explore variation and challenge the students' understanding of force (Fitzallen & Watson, 2020). For this activity, the importance of variation was reinforced in a context where expectation in the two parts of the activity was different. It was considered important to use the variation displayed in the data from the distributions for both the pre-fabricated and adjusted catapults to assist in making a decision about the "improved" performance (expectation) of the catapults.

Although sources exist online with instructions for building catapults and for lesson plans or worksheets to use in the classroom at various grade levels (e.g., Fitzgerald, 2001; Surles et al., 2011), this activity was chosen because research on statistical outcomes for elementary school students' learning associated with such activities does not appear to exist. English et al. (2013) reported working with Grade 8 students who used catapults to focus on STEM learning from an engineering perspective, which although it traced the design process developments of the students when they designed, built, and tested catapults, did not include the collection of data for statistical analysis. In research with primary classroom teachers, Jacobbe (2012) used interview questions based on prepared graphical representations of outcomes from two different catapults to illustrate the teachers' abilities to make decisions based on the shape of the distributions. The study reported here extends the work of English et al. and embraces Jacobbe's acknowledgement that data from catapult trials provide meaningful opportunities for data representation and interpretation. It also follows the same process as Petrosino et al. (2003) within an extended activity by first reinforcing variation for the initial trials before introducing an expected difference in two sets of trials.

In terms of the statistical content, the demand of the practice of statistics for context placed the activity within the realm of task design described by Ainley et al. (2006), built on the background of Realistic Mathematics Education (RME). Although RME stresses the importance of engaging students in learning contexts outside of procedurally-driven classrooms by looking outside of school, here STEM provided a legitimate context from across the school curriculum. Ainley et al. stressed the importance of avoiding superficiality, indicating that students must find both the task and the mathematics to be meaningful in the sense of their purpose and utility. In the catapult activity it was intended that students initially experience the purpose of ensuring consistency in their launches and then utilize this understanding when exploring an increase in the distance the ping pong balls travelled after changes were made. They were also introduced to the utility of measurement and technology for documenting and justifying claims about the improvements made to the catapults.

2.3. PARTICIPANTS

Two classes of Grade 4 students at an urban independent Catholic school participated in the study, with 50 students having parental permission for their data to be included in the research reported here. The sample consisted of 30 boys and 20 girls, with an average age of 10 years 2 months. The project had approval from the Tasmanian Social Sciences Human Research Ethics Committee (H0015039). For reporting purposes, students were assigned unique student identification codes (e.g., ID123) to maintain anonymity and confidentiality.

2.4. PROCEDURE

The activity took place over two sessions, five weeks apart. In the first session, the introduction to the activity included a teacher demonstration explaining the catapult and the way in which it was constructed (see Figure 3). The class discussion covered the contribution made by each part of the catapult, including the tightening of the string wrapped around the extended spoon, designated the "throwing arm." Students were familiar with catapults generally and contributed to the discussion. As well, students were reintroduced to the practice of statistics and how the activity fit the four steps: Pose a question, plan and collect data, analyze the data, and draw a conclusion (Franklin et al., 2007). The initial inquiry questions for the students were: "How far does the ping pong ball travel?" and "How consistent are we at launching the ping pong ball?" Also, class discussion took place about conducting fair tests to collect data, from which rules were established for the whole class on how the catapults would be launched.

Figure 3. The catapult used in the activity

Following the class discussion, students worked in groups of three to launch ping pong balls from a catapult and record the distance the ping pong balls travelled (Figure 4). Each student conducted four trials, with all students in the group recording the data in workbooks provided. Students were then asked to create representations of their groups' data (see Figure 5 and Fitzallen et al., 2018) to analyze their data. In their workbooks students described the consistency, variation, and reasons for the differences seen within their representations of the groups' data, followed by comparisons of their graphs with others in their groups. After a short break, students were each given a handout of their group's data (e.g., Figure 6) represented in the data analysis software *TinkerPlots* (Konold & Miller, 2015) and asked to compare it with their hand-drawn graphs. Finally, students were shown the combined data for the class (e.g., Figure 7) and asked for an overall conclusion on how far the ping pong balls travelled, the consistency of the results, and suggestions for improving the consistency. The steps of the practice of statistics were reiterated in the context of the activity undertaken.

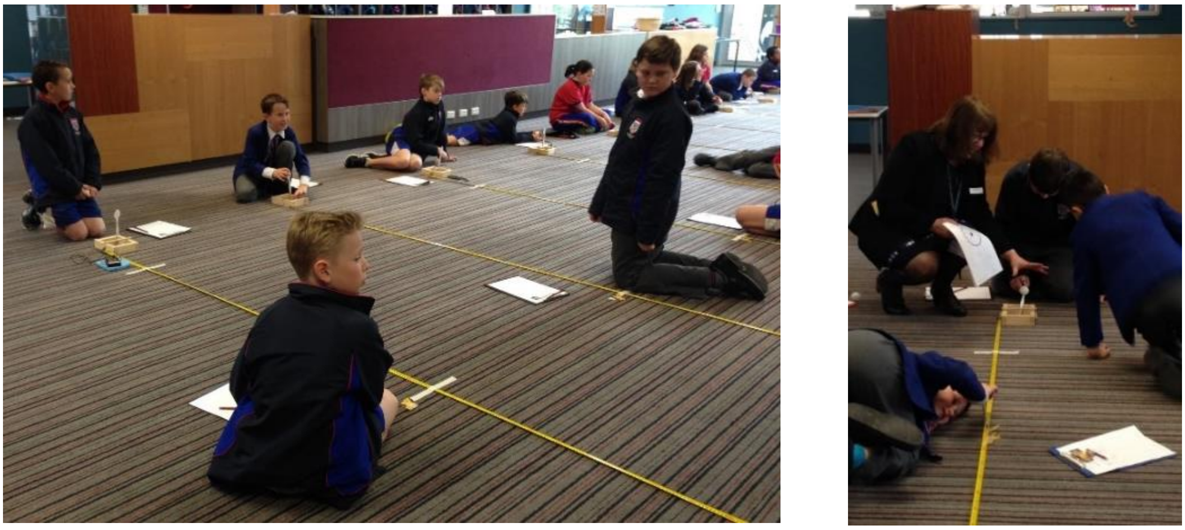
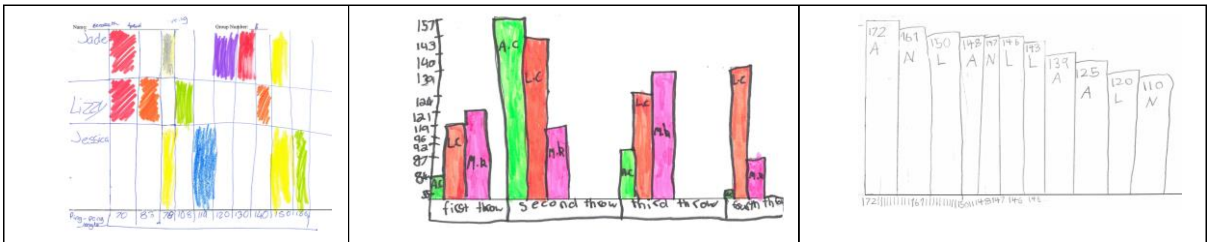


Figure 4. Trialling the catapults and recording data.



Table/pictograph | Vertical unordered value bars | Vertical ordered value bars

Figure 5. Examples of hand-drawn plots representing group data collected

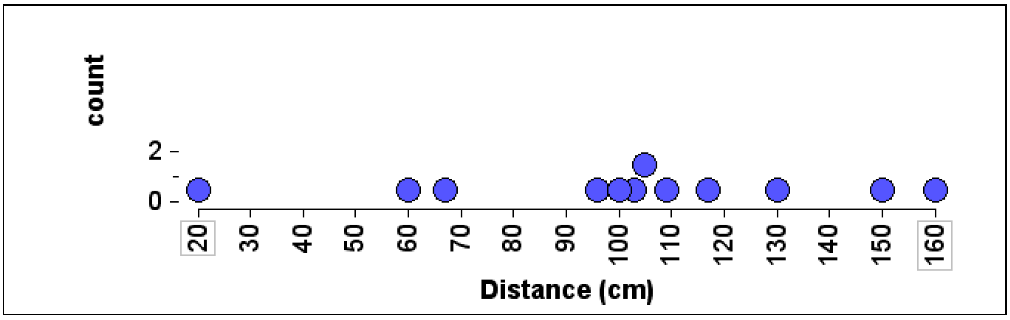


Figure 6. Example of 12 trials from one group in TinkerPlots

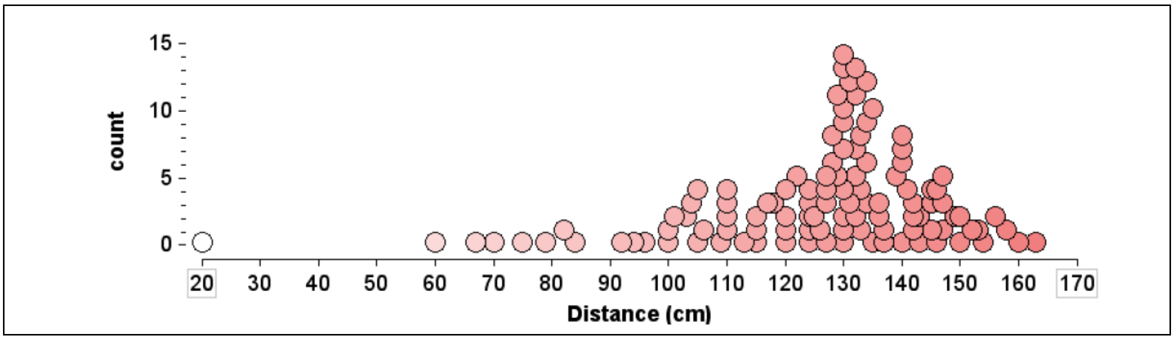


Figure 7. Data from one class for the initial catapult trials in TinkerPlots

Five weeks later, the research team returned to the classes for a second session to build on the students' experiences and to test the catapults after adjustments were made. As a starting point, the student workbooks for this part of the activity contained a plot from *TinkerPlots* of the combined class data from the previous session (e.g., Figure 7), along with review questions to answer on the range, typicality, variation, and consistency shown in the class data.

To make changes to the catapults and to look for improvement in terms of launching the ping pong balls further required two sets of trials, one with the catapults as originally set up in the first session and one with the adjusted catapults. Because of the class size and the limited time frame, possible changes to the design of the catapults were discussed, but then it was agreed that all groups would make the same change, which involved tightening the string holding the throwing arm. The research question for the students became: "How can you tell if the changes you make to the catapult improves its performance?"

After the initial trials of the catapults by groups of three with the same set up as the previous session, review questions were asked, followed by predictions for the performance of the catapults after the changes were made. After completing the second trials, students were asked to look at the data recorded in their tables and decide if the performance of the catapult had improved and how they could tell. The following day a session was held in the school's computer lab, where students used the data analysis software, *TinkerPlots*, to create plots for their groups' data, again to answer questions about the range, variation, and typical values arising from the two trials. Finally, the students were shown the combined class data and asked to describe the catapults' improved performances and how they could answer their research question from the class plots shown (e.g., Figure 8).

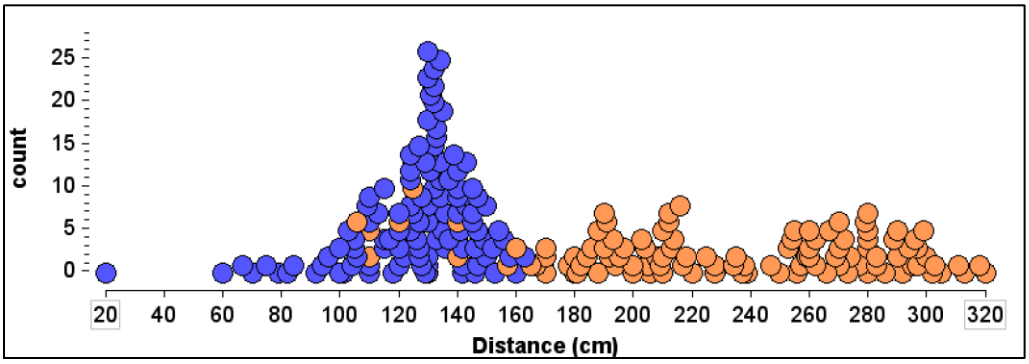


Figure 8. Class data for comparing original and "improved" catapults in TinkerPlots

2.5. DATA COLLECTION

The data collected for the current analysis of students' developing understanding of variation in the STEM activity consisted of the written responses to selected questions in the two student workbooks, and *TinkerPlots* files produced by the students. The workbook questions selected for analysis were related specifically to variation and the representations created in *TinkerPlots* to investigate the difference between the results for the two trials of the catapults. All workbook questions used in the activity are found in the Appendix, condensed in space. The questions used in this analysis were asked (a) during Part 1 of the activity for the initial testing of the catapults based on each group's data (A4, A5, and A6), (b) as a review of the graph of the class data from Part 1 at the beginning of Part 2 (B3), (c) for considering the results of each group's second trials (B10/B11 and B13), and (d) to reach a conclusion for the class data at the end of Part 2 (B16). Also included in the analysis were the *TinkerPlots* graphs created by the students.

Although the students worked together in groups of three in the school hall to collect and record their data during the trials, their responses to the workbook questions analyzed here were written by students separately in the classroom. There was no evidence of students copying answers from each other as this type of work was common in the school, and expectations were well-known by students.

2.6. DATA CODING AND ANALYSIS

To answer the study's research question, the workbook responses related to variation completed after students conducted each of the three sets of trials (one in the first session and two in the second session) were analyzed using the SOLO model of Biggs and Collis (1982, 1991), as presented in Figure 2. As well, the model was used to analyze the graphical representations that the students produced in *TinkerPlots* to determine the performance of the catapults after improvements were made.

The coding of responses to workbook questions focused on variation and the student generated *TinkerPlots* representations was carried out using the coding method shown in Table 1. The coding mirrored the three levels of the modes of the SOLO model (Biggs & Collis, 1982, 1991) extended for the IK mode as suggested by Groth et al. (2021). There are U, M, or R level IK responses that are potentially normative inconsistent with the task or normative consistent with the task (Figure 2). Similarly, there are potentially the three levels for CS responses to the tasks. The horizontal arrows represent the potential Multimodal support from IK normative consistent contributions to a final CS response. Students may hence travel vertically (in Figure 2) in an IK path, travel vertically in a CS path, or may include input (horizontally in Figure 2) from a normative consistent IK response supporting a CS response. Specifically, for the *TinkerPlots* graphs, coding associated with U, M, and R levels in the CS mode depends on whether a single variable is displayed (U), two variables are displayed using "bins" (M), or two variables are displayed separating data on a horizontal axis. Graphs not showing data are coded as IK.

In reporting the Multimodal responses, there were nine possibilities for responses, with each IK level possibly supporting each CS level. Coding was initially carried out separately by the first and third authors with constant reference to the combining of elements in students' explanations, within and across modes. Their agreement was 70%, hence they checked all responses together, mainly in confirming the U, M, or R levels of response. Because the aims reported here relate to the practice of statistics rather than the detailed description of development of a science concept, informal comments made about the physical experiences of the activity were considered to be contextual intuitions and images in the IK mode (Collis & Romberg, 1991). The normative compatibility or incompatibility of the IK responses (Groth et al., 2021) was considered during coding, but given the ages and experiences of the students, it was not expected there would be very many incompatible responses. The assumption in Table 1 and later tables is that reported IK responses are normative compatible.

Table 1. Coding for workbook responses based on Figure 2

Ikonic Mode (IK)		Levels
A qualitative description based on imagery, intuition, belief, perhaps in relation to an observation or activity	IK_U	One such observation
	IK_M	Two or more such observations in a list
	IK_R	Two or more such observations linked in an explanation
Concrete Symbolic Mode (CS)		Levels
A quantitatively-based reference, including numbers, calculations, measurements, and symbolic devices	CS_U	One such reference
A quantitatively-based reference, including numbers, calculations, measurements, and symbolic devices	CS_M	Two or more such individual references
	CS_R	Two or more such references linked in an explanation
IK-CS Multimodal responses		Levels
Responses including one of the three levels of IK support at the three CS levels	IK-CS_U	IK_U, IK_M, or IK_R observation/s in support of a single CS reference
	IK-CS_M	IK_U, IK_M, or IK_R observation/s in support of two or more CS quantitative references in a list
	IK-CS_R	IK_U, IK_M, or IK_R observation/s in support of two or more CS references linked in an explanation

3. RESULTS

The tabular format for reporting student understanding of variation displayed during the activity presents first the responses in a single mode, either IK or CS, followed by the results for the Multimodal responses with IK support. This is done first for the Session 1 trials of the catapuplts, second for the Session 2 review of the first trials, and third for the Session 3 determination of the result of the changes made. The analysis of student workbook responses revealed only two IK responses that were considered normative incompatible. These responses are noted in the following presentation, but otherwise it should be assumed that all IK responses are normative compatible. Following the results for the workbook questions and the *TinkerPlots* graphs, a summary is presented of the SOLO levels across the eight responses presented.

3.1. SESSION 1: INITIAL TRIAL

Using the representations drawn from the data collected from the initial trials of the catapults (see Figure 5 and Fitzallen et al., 2018), students were asked in their workbooks to interpret the data in terms of how consistent the data (A4) were and to give examples of variation (A5). Indicative examples from the two sets of written responses are shown in Tables 2 and 3. (Student ID numbers are provided for each response). Because of the hands-on experience of the physical context and its visual representation in the graphs produced, most of the responses included language that was qualitative in nature (IK), reflecting understanding of the context. For some students (e.g., 42% for Question A4 and 38% for Questions A5), this was their complete response. For a similar proportion of students (e.g., 46% for A4 and 48% for A5), however, it was the foundation for also reporting numerical values to support the claims made (Multimodal). In all cases, the students' misinterpretation of the question related to the response discussing, in general, the graphs produced by different students with a group, either agreeing with each other (A4) or being different (A5 and A6), without actually considering the variation of the data within the graphs.

Table 2. Responses related to consistency seen in the data

Level	Question A4: How consistent were your data?*	%
IK_U	Not that consistent. [ID115]	12%
IK_M	It was very consistent but a few were very different. [ID121]	8%
IK_R	Quite consistent, because we had most of our landings very close together, practically all bunched together. [ID145]	22%
CS_U		0%
CS_M	Some were about 30 cm apart and others were sort of 150 to 170. We had a period of time where everyone was getting about the 160 mark. [ID152]	4%
Multimodal		%
$IK\text{-}CS_U$	IK_U: Most of the balls in our team went over 60. Pretty consistent. [ID101] (n = 5)	26%
	IK_M: It wasn't really consistent but some of them were similar for example 122 and 120. [ID114] (n = 6)	
	IK_R: It was very consistent because it was really only around 140's to 180's and we pulled the spoon down as far as we could for most of it [ID102] (n = 1)	
$IK\text{-}CS_M$	IK_U: Pretty consistent but there were about 3 numbers that weren't very close. The main areas of numbers were 120 to 170. [ID112] (n = 5)	18%
	IK_M: Our lowest 55 and our highest which was 157. There were little groups of tape together and there was only 1 lonely tape which was 55. [ID158] (n = 4)	
$IK\text{-}CS_R$	IK_M: It was not very consistent. My highest was 139 and my lowest was 87. The two others in the middle of 139 and 87 were 121 and 96. The one on it's own is number 55. [ID117] (n = 1)	2%

* Four students misinterpreted the question.

Table 3. Responses related to variation and differences seen in the data

Level	Question A5: What examples of variation are there in the data?*	%
IK_U	There was only 2 or 3 variations so very little variation. [ID128]	16%
IK_M	Joseph got the least and got the most. The graph is going up and down. [ID103]	10%
IK_R	Our biggest one and our smallest one are very far apart but all the others are very close together. [ID116]	12%
CS_U	There was only one four and that was the highest number. [ID139]	2%
CS_M	Lowest 107 cm. Highest 148 cm. Variation 31 cm from highest to lowest. [ID140]	6%
Multimodal		
$IK\text{-}CS_U$	IK_U: There was a big difference in the numbers because it was from 139 down to 87. [ID117] (n = 1)	8%
	IK_M: 74 was the furthest and the rest were kind of bunched up. [ID121] (n = 1)	
	IK_R: Some had a massive gap up to 30 cm and there were some short ones and some long ones, there was tonnes of variety. [ID152] (n = 2)	
$IK\text{-}CS_M$	IK_U: Some were only around the 120's mark and some were around the 160's mark. There were big differences. [ID149] (n = 10)	40%
	IK_M: The first shot we did was the longest, 172 cm and our last shot was the smallest, which was 110 cm. [ID145] (n = 10)	

* Two students misinterpreted the question, and one did not answer.

Explanations in the workbooks about the differences among the data from the individual throws (A6) were judged to be IK responses, with only one considered normative incompatible with the context of the activity: "Luck I think that some of it was from luck" [ID136]. See Table 4 for exemplary responses.

Table 4. Conclusions about the data from the first session

Level	Question A6: Why might there be differences in the data?*	%
IK_U	Because maybe we didn't pull the spoon down far enough than the others we did so it didn't go that far. [ID102]	46%
	Some of us are stronger. [ID134]	
IK_M	I think maybe because it was different people and maybe one of us did not pull back as far as the other person did. [ID145]	34%
	Because sometimes the catapult went wobbly and some shots went higher than longer. We all pulled the spoon back full length but it still went way different. [ID152]	
IK_R	Maybe some people pressed harder to make it go forward. But if it flies up the ball goes down. Some people pressed gently and it went forward. So usually there's a difference because about the press. [ID104]	6%

* One student misinterpreted the question, and six did not answer.

3.2. SESSION 2: REVIEW OF FIRST TRIAL WITH ORIGINAL SETUP

At the beginning of the second session to explore improving the performance of the catapults, students were asked in their workbooks to review the data for their class of the trials from the previous session (B3; e.g., Figure 7). The responses in Table 5 illustrate the broad range of responses elicited. The responses included the different modes with Multimodal functioning displayed in 58% of responses, and the rest split evenly between IK and CS modes. Again, describing the visualization added support to the validity of the numerical data reported.

Table 5. Review of class data from the first trial at the start of the second session

Level	Question B3: Where do you see variation in the plot? Give an example. (see Figure 7)*	%
IK_U	Difference. There is very big variation to the first couple near the end. [ID130]	12%
IK_M	One was almost at the bottom and one was almost at the top. [ID116]	6%
IK_R	At the first half of the plot there is a lot of variety but in the second half everything started to get even. [ID152]	2%
CS_U	50 and less [ID125]	8%
CS_M	Between 4 and around 57 which is about 53 spaces. [ID117]	12%
Multimodal		
IK-CS_U	IK_U: At the start 10 cm to 20 cm to 30 cm the rest is bunched up. [ID111] (n = 6)	14%
	IK_M: There was a lot of variation between 10 cm to 109 cm because there was not much stacking so most of them were different. [ID142] (n = 1)	
IK-CS_M	IK_U: I see it everywhere especially between 10 and 190. That's a 180 cm variation. [ID113] (n = 12)	42%
	IK_M: 10 and 140 cm because on 10 there's barely any catapult launches and 140 because it's all stacked up. [ID144] (n = 9)	
IK-CS_R	IK_U: Inbetween 10 and 50 are not big compared to inbetween 130–170. [148] (n = 1)	2%

* One student did not answer the question.

Analysis of this workbook question about variation in the plot included the second recorded normative incompatible comment: "You see a variation where one ball reached 10 cm and another ball on 190 cm like maybe someone pulled one too far back to get 10 cm. Maybe it was a lucky shot to get 190 cm." [ID155]. The rest of the IK support was considered normative compatible with the context, and the complete response was coded as IK_M-CS_M.

3.3. SESSION 2: DETERMINING THE EFFECT OF CHANGES MADE

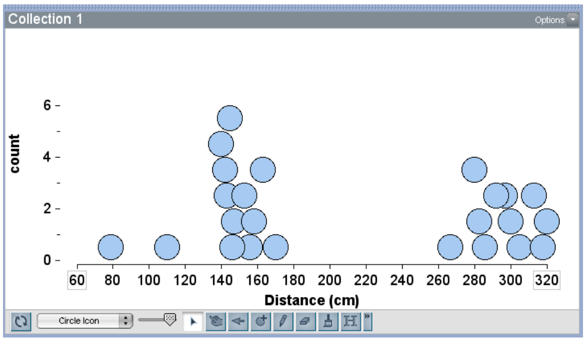
Analysis of the workbook responses made when asked about the data collected from the trials after again testing and then improving the catapults (B10/B11) revealed that 38% of responses reflected descriptive (IK) accounts of the outcomes, and 40% of responses reflected the numerical (CS) information available, with 18% including IK support for a CS response (Multimodal functioning, see Table 6).

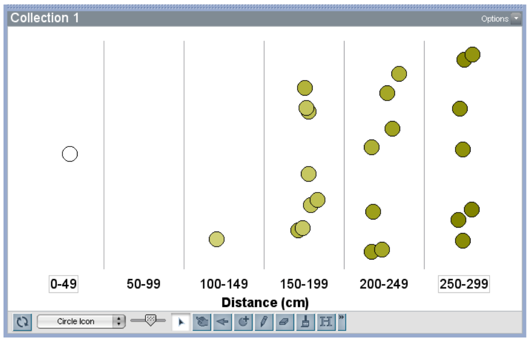
Table 6. Improvement after the catapults were changed

Level	Question B10/B11: How can you tell from the data you collected that the performance of the catapult improved?*	%
IK_U	Because the ping pong balls went much further than last time. [ID112]	16%
IK_M	Because it has gone over the tape measure and sometimes go to the end of the tape measure. [ID140]	10%
IK_R	Because it was tighter so the catapult gave the ball more energy so the ball would go further. [ID135]	12%
CS_U	Well nearly all of them are in the two hundreds. [ID155]	8%
CS_M	The highest was 154 the first time and the highest this time was 311. [153]	6%
CS_R	That on the measurer on the first data page the data was further down the line but in the 2nd it was further up the line. [ID117]	26%
Multimodal		
IK-CS_U	IK_U: It is more consistent most of the throws are over 200. [ID149] (n = 2)	4%
IK-CS_M	IK_M: We got 80 that was our lowest and 395 there was more consistents [*sic*] than variation. [ID119] (n = 2)	4%
IK-CS_R	IK_M: Because it was a bigger length than before because our lowest was 20 and now our highest is 302. [ID144] (n = 5)	10%

* Two students did not answer the question.

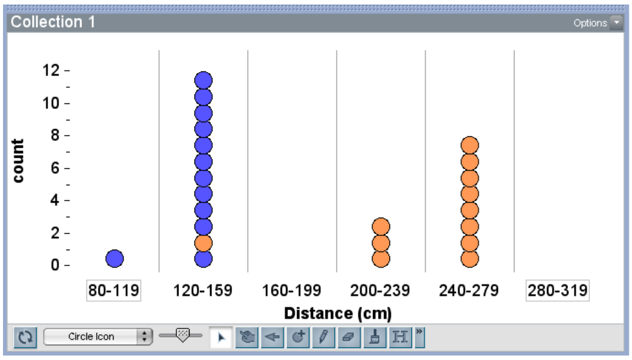
Forty-seven of the 50 students saved at least one plot in *TinkerPlots* during the session in the computer lab. For students who created more than one plot, the one that conveyed the most information about the variation between the two trials was selected for analysis. Five categories of plots are shown in Figure 9, accounting for 45 of the plots created. Of the two not featured in Figure 9, one was an unlabelled pie graph, and the other consisted of lines marked with the pencil over a random display of the data. These were considered normative compatible IK_U representations as they were visual but did not engage with the CS numerical aspects of the data. Ten of the plots (20%) represented a single variable (CS_U), the distance travelled, without distinguishing the two trials. Seven representations (14%) distinguished the two variables in categories (either by color or by using two axes) that demonstrated the difference in the two trials (CS_M), whereas 28 (56%) went further to show the actual values obtained for each trial (CS_R).

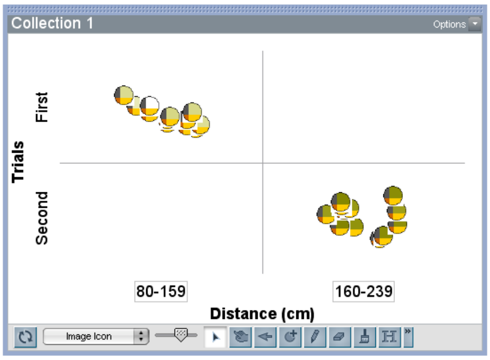




Single variable without distinction for trial (n = 10)

CS Unistructural (20%)

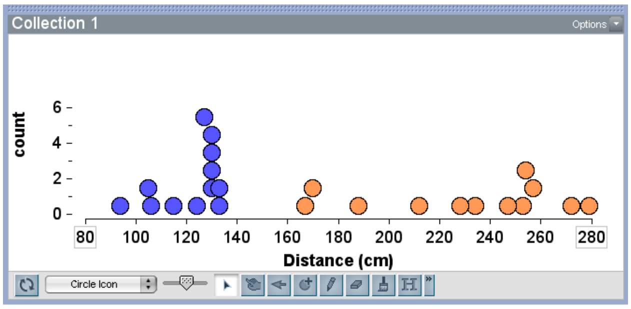


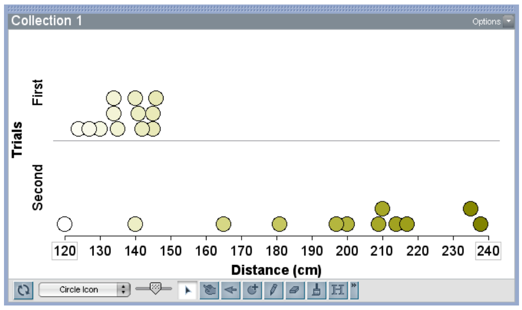


Single axis bins (color difference) (n = 3)

Split axes bins (n = 4)

CS Multistructural (14%)





Single axis (color difference) (n = 13)

Split axes by trials (n = 15)

CS Relational (56%)

Figure 9. Representations created in TinkerPlots by the students

When asked for the ranges for the data in the two trials (B12), 96% of students could report these values. Asking for a description of the variation seen in the data (B13) was more difficult. Table 7 illustrates the extent to which students could describe in their workbooks the types of variation now visible in their representations of the data. The use of the word “describe” in the question may have encouraged more IK responses to this question (60%). Asked specifically for the typical distances travelled in the two trials (B14), 68% gave reasonable responses for each. “Reasonable” had to be judged separately for each response because the plots created were all potentially different for the data for different groups. The decisions were agreed on by the two authors doing the coding.

Table 7. Variation in the second trials for group data

Level	Question B13: Describe the variation in your group's data.*	%
IK_U	Trial 2 goes way way further than trial 1. [ID158]	22%
IK_M	There is bigger gaps in the second trial than the first trial. Because trial 1 is more lumped up than the second trial. [ID114]	28%
IK_R	The first test was very consistent. But the second test had a lot of variation because some people could have terned [turned] it to[o] far or not anuf [enough] and people copet [?kept] pulling it to[o] far back. [ID101]	10%
CS_U	29–82 [ID115]	2%
CS_M	There's variation between 203 cm and 296 cm. [ID138]	10%
Multimodal		
IK-CS_U	IK_U: The variation in our tinker plots were in the first trial because 82 was far from the rest of the bunch. [ID149] (*n* = 2)	4%
IK-CS_M	IK_U: In the first data trials there is only a little bit of variation between 138 and 157. [ID119] (*n* = 1)	14%
	IK_M: From 73 to 100 there is a massive gap and from 203 to 296 there is another big gap. In plot 2 there is a big gap between 159 to 220 and 220 to 265 is another massive gap. [ID125] (*n* = 5)	
	IK_R: In the 1[st] trial it did not have much variation, our best throw travelled 133 cm, our worst throw was 70 cm long, that happened because Jesse pulled the spoon back really far, actually way too far. [ID123] (*n* = 1)	
IK-CS_R	IK_U: Tightening the string made a big, big variation in the distance of the ball being flung. It was a distance between of 44, between 1st's highest and 2nd's lowest. [ID117] (*n* = 1)	6%
	IK_M: In the first trials it was all below 160 but in the second trial every throw was over 160. In our first trials it was very consistent, most of our throws were in the 130's. But in our second trials it was everywhere. [ID152] (*n* = 1)	
	IK_R: There wasn't much variation in the first one because the dots were clumped together in 94 and 133. But the second trial had a little bit of variation because they were spread out in 167 and 279. [ID102] (*n* = 1)	

* Two students did not answer the question.

Finally, when viewing the class plot (e.g., Figure 8), students were asked if changing the number of turns had improved the performance of the catapults (B15). All who responded said "yes". When specifically asked how they could determine this, only one did not use IK language in justification of an answer. Forty-four responses were solely IK in nature, whereas the remaining four responses were Multimodal in adding CS justification from the plot (Table 8).

Table 8. Final conclusion from the class data

Level	Question B16: How can you tell from the plot if the catapults have been improved or not?*	%
IK_U	It has improved by heaps. [ID115]	18%
IK_M	In the second one it was more spread out and it went further. [ID130]	24%
IK_R	Because in the first throws [we] were a bit behind but when we turned the string handles, because it got tighter it flung much more harder, so it went further. So on the screen the orange dots were further." [ID102]	46%
CS_R	I can tell it improved because we had it closer to zero but now the furthest is 484 cm. [ID105]	2%
Multimodal		
$IK\text{-}CS_M$	IK_R: Number 1's highest is 113 and number 2's highest is 213. Pulling the strings made the catapult go further. [ID111]	2%
$IK\text{-}CS_R$	IK_R: It improved because the 1st trial was only 150 cm and the second trial was 420 cm. Because we did 6 half turns on the 2nd trial and we didn't do any half turns at the 1st trial." [ID141]	6%

* One student did not answer the question.

3.4. SUMMARY OF RESPONSE LEVELS ACROSS QUESTIONS

There were considerable differences in the levels and modes of cognition shown across the different tasks analyzed in this report (Tables 2 to 8). These are summarized for the three possible solution pathways in Table 9 for the students who responded to each of the questions. The high percentages of IK responses reflect the appreciation of the physical context in considering variation in an environment where students spent a considerable amount of hands-on time collecting, checking, recording, and graphically representing numerical data. As well, in many cases there appeared to be support of CS responses from appropriate IK observations. In some cases, the qualitative descriptive language of the IK mode appeared to prompt the necessity to provide further CS numerical evidence of an answer (e.g., ID114 for Question A4, ID145 for A5, ID113 for B3), or subsequently, to justify descriptively (IK) the numbers provided (CS) (e.g., ID158 for A4, ID121 for A5, ID111 for B3). Although this decreased after the completion of the second data collection, in five out of six workbook questions there were more CS responses with Multimodal support than without.

Table 9. Summary of percentage of SOLO responses across the activity

Question	IK	CS	Multimodal
A4: How consistent were your data?	46%	4%	50%
A5: What examples of variation are there in the data?	40.4%	8.5%	51.1%
A6: Why might there be differences in the data?	100%		
B3: Where do you see variation in the plot? Give an example.	20.4%	20.4%	59.2%
B10/11: How can you tell from the data you collected that the performance of the catapult improved?	39.6%	41.7%	18.7%
TinkerPlots representation (Figure 9).	4%	96%	
B13: Describe the variation in your group's data.	62.5%	12.5%	25%
B16: How can you tell from the plot if the catapults have been improved or not?	90%	2%	8%

4. DISCUSSION

4.1. CHARACTERIZING THE UNDERSTANDING OF VARIATION WITHIN A STEM INQUIRY

This research takes heed of Green's (1993) and Shaughnessy's (1997) pleas for the study of what students "*are* capable of doing" when experiencing variation. Exploring students' explanations of their experiences of variation in a context was accomplished by building on the recent developmental work of Groth et al. (2021) in the Ikonic mode (IK) of functioning, as well as further exploring the Concrete Symbolic mode (CS) of functioning (Biggs & Collis, 1982, 1989) and the relationship of the two via Multimodal functioning (Collis & Romberg, 1991; Watson et al., 1995; Watson & Collis, 1994). The link between the normative compatible IK mode and the CS mode is seen to provide considerable support for many of the CS responses, as shown in the detailed analyses of the responses to workbook questions and the summary in Table 9.

This study has shown that looking in more detail at the contextual thinking about variation in the IK mode, as suggested by Groth et al. (2021), can be useful in appreciating how students prepare to take on the demands of the CS mode in a particular context. The Groth et al. study of the IK mode did not include links to the CS mode. The activity that was the context for this study allowed this extension to occur and suggested that there appears to be many instances where IK thinking prompts or supports a response in the CS mode. The comment of Biggs and Collis (1991) concerning CS mode reasoning not replacing IK mode reasoning but there being a potentially positive relationship between them appears to be supported in this study. Analysis of the workbook responses revealed there was facilitation of the CS understanding from the IK mode, particularly in Tables 2 and 4, where there were more CS responses with IK support than without IK support.

It is now interesting to consider again the research of Reading (2004) on students' understandings of variation, in suggesting two hierarchical cycles of response in the CS mode, one qualitative and one quantitative. The qualitative responses were related to the context of the study but did not employ numerical information as would reflect the CS mode. Perhaps instead of considering the refined two-stage CS hierarchy developed in that analysis, the data could be considered a strong case of Multimodal functioning employing the use of contextual knowledge as IK support in a context where it is highly relevant.

These findings supplement the foundational research conducted by Watson et al. (2003) on understanding variation across a range of contexts based on Rasch analysis of surveys. The use of the SOLO model provides additional structure to the four levels of reasoning observed there. For example, the U-M-R structure of the IK mode can further elucidate the Prerequisites level of intuitive thinking about variation in the earlier study. In considering the original educational application of the SOLO model's five modes (Biggs & Collis, 1982), they were seen as hierarchical, with the aim to move students to increasingly higher modes of operation, particularly in the realms of pure mathematics. In the area of statistics where context is necessary, however, and in the primary years, where the CS mode is dominant, it appears that IK support, particularly if compatible with the context of the task, can be helpful in achieving CS responses.

The much-cited claim by Moore (1990) that the study of statistics needs to be embedded within meaningful contexts holds true for this study. This study, however, goes further to show that there are multiple contextual aspects of an investigation of variation that influence sense-making of the data and decision-making from the data. The overall context of this STEM inquiry was related to determining if the performance of a catapult was improved (that is, varied) after adjustments were made to the firing mechanism. In making determinations about the performance of the catapults the students drew on the multiple aspects of variation in the data collection context (launching the ping pong balls from the catapults) and in the context of analyzing the collected data (drawing on the affordances of *TinkerPlots* to make sense of the variation) to answer their research question. It is the coming together of the thinking about these three differing contextual aspects that makes the Catapult activity valuable across the STEM disciplines.

4.2. FURTHER THOUGHTS

Reflecting on the results of this study, the way questions are posed for students may influence the inclusion of IK components in answers. Only for the workbook question specifically referring to data, "How can you tell from the data you collected that the performance of the catapult improved?" (Table 6) were more than 40% of the actual responses in the CS mode without IK support. For the other questions, including words such as "consistency" and "variation", the percentage of CS-only responses ranged from 4% to 20%. In recalling that this was considered a STEM activity, it appears reasonable, and encouraging, that the context of carrying out the study was a factor encouraging IK support for CS responses ranging from 25% to 59% (Tables 2, 3, 5, and 7). The aim as students move across the years is to link more mathematical measures to their ability to describe variation, for example, the range, mean absolute difference, or standard deviation (Kader & Mamer, 2008). In most contexts, however, it will still remain valuable to gain clues from hands-on experiences of variation that support a numerical value provided or mathematical description in the CS mode.

Further, looking at the data from the perspective of the students' responses rather than the questions, again for the five questions in Tables 2, 3, 5, 6, and 7 and the three general categories of IK, CS, and Multimodal, five students responded with only one type of response for all questions. Four of these gave all IK responses and one all Multimodal responses. Seventeen students provided responses of all three types, whereas 28 students gave responses of two types; of these, 17 were IK and Multimodal, eight were IK and CS, and three were CS and Multimodal. Hence all students except three gave at least one IK response, 28 gave at least one CS response, and 38 gave at least one Multimodal response. This would appear to encourage IK contributions and support in the early activities for students in STEM topics where the context can contribute to higher level thinking, as well as being a starting point from hands-on observations.

5. CONCLUSION

Much research has been carried out employing the SOLO model to determine the development of understanding of statistical concepts (e.g., Watson et al., 2003; Watson & Kelly, 2005; Watson et al., 2020b). The application of the U-M-R levels to the IK, CS, and Multimodal responses in this study goes further to illustrate that student thinking about the context and the variation within the context may also be potentially helpful in building understanding not only of variation but also of force and energy. Employing the multi-level aspect of the SOLO model in analyzing the workbook responses was useful in the STEM environment of this study because of the recognition of the influence of the context on the responses. Given that STEM activities involving data collection will provide meaningful real-world contexts involving variation and that data are only meaningful when arising from context (Cobb & Moore, 1997), analyzing responses in a manner that recognizes the IK contribution to the CS aspects of learning about the practice of statistics makes logical sense. Further research in other STEM contexts should provide opportunities to explore further this relationship in more detail.

As noted by Watson et al. (2020a), statistics presents one of the fastest growing job opportunities in the STEM employment field, as well as more generally across society. Starting early, with the appreciation that the interaction of STEM contexts and intuitive statistical ideas can support developing understanding of the fundamental concepts of statistics, gives promise for motivating students' interest in problem-solving across the STEM fields. Although variation was the focal concept of the practice of statistics in this report, three other fundamental concepts were also supported across the activity: distribution, through the movement from hand-drawn to *TinkerPlots* representations; expectation, through the "hypothesis" that the catapults could be improved to launch ping pong balls further; and informal inference, through providing supporting evidence that indeed improvement had occurred.

Based on this STEM-based topic and others (e.g., Fitzallen & Watson, 2020), it seems plausible to say that it is possible to provide learning opportunities for students to build understanding of variation and the practice of statistics in STEM contexts that facilitate developing meaningful decision-making skills. That no other report of a catapult activity including data collection at the school level was found before starting the research was a surprise to the authors. The hope is that this study will convince teachers and researchers across the compulsory years of schooling that there are meaningful and useful opportunities to complement science, and other STEM activities, with the practice of statistics.

ACKNOWLEDGEMENTS

This study was funded as part of an Australian Research Council (ARC) Discovery Grant (DP150100120). Any opinions, findings, conclusions, or recommendations reported are those of the authors and do not necessarily reflect the views of the ARC. Jane, Noleine, and Ben acknowledge the assistance of Bruce Duncan, who designed the catapult and introduced the device to the students, and express extreme sadness at the untimely passing of their co-author Suzie Wright.

REFERENCES

Ainley, J., Pratt, D., & Hansen, A. (2006). Connecting engagement and focus in pedagogic task design. *British Educational Research Journal*, *32*(1), 23–38. https://doi.org/10.1080/01411920500401971

Australian Curriculum, Assessment and Reporting Authority. (2019). *Australian Curriculum*. https://www.australiancurriculum.edu.au/

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 (GAISE II)*. American Statistical Association and National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press.

Biggs, J., & Collis, K. (1989). Towards a model of school-based curriculum development and assessment using the SOLO taxonomy. *Australian Journal of Education*, *33*(2), 151–163. https://doi.org/10.1177%2F168781408903300205

Biggs, J. B., & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behaviour. In H. A. H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 57–76). Lawrence Erlbaum.

Chick, H., Watson, J., & Fitzallen, N. (2018). “Plot 1 is all spread out and Plot 2 is all squished together”: Exemplifying statistical variation with young students. In J. Hunter, P. Perger, & L. Darragh (Eds.), *Making waves, opening spaces*. Proceedings of the 41st annual conference of the Mathematics Education Research Group of Australasia, Auckland, July 1–5 (pp. 218–225). https://merga.net.au/Public/Publications/Annual_Conference_Proceedings/2018-MERGA-CP.aspx

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *American Mathematical Monthly*, *104*, 801–823. https://doi.org/10.1080/00029890.1997.11990723

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13. https://doi.org/10.3102%2F0013189X032001009

Collis, K.F., & Romberg, T.A. (1991). Assessment of mathematical performance: An analysis of open-ended test items. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition*. Prentice Hall.

Creswell, J. W. (2013). *Research design: Quantitative, qualitative and mixed method approaches* (2nd ed.). SAGE Publications.

Engler, J. (2012, June 15). STEM education is the key to the U.S.’s economic future. *U.S. News and World Report Civic*. https://www.usnews.com/opinion/articles/2012/06/15/stem-education-is-the-key-to-the-uss-economic-future

English, L. D., Hudson, B., & Dawes, L. (2013). Engineering based problem solving in the middle school: Design and construction with simple machines. *Journal of Pre-college Engineering Education Research*, *3*(2), 1–13. https://doi.org/10.7771/2157-9288.1081

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education*, *7*(2). http://escholarship.org/uc/item/7gv0q9dc#page-1

Fitzallen, N., & Watson, J. (2020). Using the practice of statistics to design students’ experiences in STEM. In B. Shelley, K. te Riele, N. Brown, & T. Crellin (Eds.), *Harnessing the transformative power of education* (pp. 74–99). Koninklijke Brill. http://dx.doi.org/10.1163/9789004417311

Fitzallen, N., Watson, J., Wright, S., & Duncan, B. (2018). Data representation in a STEM context: The performance of catapults. In M. A. Sorto & E. Paparistodemou (Eds.), *Looking back, looking* *forward*. *Proceedings of the 10th International Conference on the Teaching of Statistics*, Kyoto, Japan, July 8–14. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_4B2.pdf?1531364264

Fitzgerald, M. (2001). Katapultos: Teaching basic statistics with ballistics. *Tech Directions*, *61*(4), 20–23.

François, K., & Monteiro, C. (2018). Big data literacy. In M. A. Sorto & E. Paparistodemou (Eds.), *Looking back, looking forward. Proceedings of the 10th International Conference on the Teaching of Statistics*, Kyoto, Japan, July 8–14. International Statistical Institute. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_7F1.pdf?1531364287

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. http://www.amstat.org/education/gaise/

Green, D. (1993). Data analysis: What research do we need? In L. Pereira-Mendoza (Ed.), *Introducing data analysis in the schools: Who should teach it?* (pp. 219-239). International Statistical Institute.

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*, 241-261. https://doi.org/10.1007/s13394-019-00287-w

Jacobbe, T. (2012). Elementary school teachers' understanding of the mean and median. *International Journal of Science and Mathematics Education, 10*, 1143–1161. https://doi.org/10.1007/s10763-011-9321-0

Kader, G., & Mamer, J. (2008). Statistics in the middle grades: Understanding center and spread. *Teaching Mathematics in the Middle School*, *14*(1), 38–43. https://doi.org/10.5951/MTMS.14.1.0038

Konold, C., & Miller, C. D. (2015). *TinkerPlots: Dynamic data exploration.* [Computer software, Version 2.3.2]. Learn Troop.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82–105. http://iase-web.org/documents/SERJ/SERJ8(1)_Makar_Rubin.pdf

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95–137). National Academy Press. https://doi.org/10.17226/1532

Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics*. W. H. Freeman.

National Research Council. (2013). *Next generation science standards: For states, by states*. National Academy Press. https://doi.org/10.17226/18290

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

Office of the Chief Scientist. (2013). *Science, technology, engineering and mathematics in the national interest: A strategic approach*. Australian Government. https://www.chiefscientist.gov.au/sites/default/files/STEMstrategy290713FINALweb.pdf

Petrosino, A. J., Lehrer, R., & Schauble, L. (2003). Structuring error and experimental variation as distribution in the fourth grade. *Mathematical Thinking and Learning, 5*(2&3), 131–156. https://doi.org/10.1080/10986065.2003.9679997

Reading, C. (2004). Student description of variation while working with weather data. *Statistical Education Research Journal*, *3*(2), 84–105. https://iase-web.org/documents/SERJ/SERJ3(2)_Reading.pdf?1402525005

Ridgway, J., Ridgway, R., & Nicholson, J. (2018). Data Science for all: A stroll in the foothills. In M. A. Sorto & E. Paparistodemou (Eds.), *Looking back, looking forward. Proceedings of the 10th International Conference on the Teaching of Statistics*, Kyoto, Japan, July 8–14. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_3A1.pdf?1531364253

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph & K. Carr (Eds.), *People in mathematics education*. *Proceedings of the 20th annual Conference of the Mathematics Education Research Group of Australasia, Aotearoa* (Vol. 1, pp. 6–22). https://www2.merga.net.au/documents/Keynote_Shaughnessy_1997.pdf

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading, C. (1999, April). School mathematics students' acknowledgment of statistical variation. In C. Maher (Chair), *There's more to life than* *centers*. Presession Research Symposium, 77th Annual National Council of Teachers of Mathematics Conference, San Francisco, CA.

Surles, W., Crosby, J., McNeil, J., Zarske, M. S., & Samson, C. (2011). *Launch into learning: Catapults!* College of Engineering and Applied Science, University of Colorado. https://www.teachengineering.org/lessons/view/cub_catapult_lesson01

Watson, J. M. (2005). Variation and expectation as foundations for the chance and data curriculum. In P. Clarkson, A. Downton, D. Gronn, M. Horne, A. McDonough, R. Pierce & A. Roche (Eds.), *Building connections: Theory, research and practice*. *Proceedings of the 28th Annual Conference of the Mathematics Education Research Group of Australasia,* Melbourne (pp. 35–42). MERGA.

Watson, J.M. (2007). Inference as prediction. *Australian Mathematics Teacher*, *63*(1), 6–11.

Watson, J.M. (2009). The development of statistical understanding at the elementary school level. *Mediterranean Journal for Research in Mathematics Education*, *8*(1), 89–109.

Watson, J.M., & Callingham, R.A. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3–46. http://iase-web.org/documents/SERJ/SERJ2(2)_Watson_Callingham.pdf

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

Watson, J. M., Campbell, K. J., & Collis, K. F. (1993). Multimodal functioning in understanding fractions. *Journal of Mathematical Behavior*, *12*(1), 45–62.

Watson, J. M., & Collis, K. F. (1994). Multimodal functioning in understanding chance and data concepts. In J. P. da Ponte & J. F. Matos (Eds.), *Proceedings of the Eighteenth International Conference for the Psychology of Mathematics Education*, Lisbon (Vol. 4, pp. 369–376).

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.M., Collis, K.F., & Moritz, J.B. (1997). The development of chance measurement. *Mathematics Education Research Journal*, *9*, 60–82. http://dx.doi.org/10.1007/BF03217125

Watson, J., Fitzallen, N., & Chick, H. (2020a). What is the role of statistics in integrating STEM education? In J. Anderson & Y. Li (Eds.), *Integrated approaches to STEM education: An international perspective* (pp. 91–116). Springer. https://doi.org/10.26181/601cd997becd5

Watson, J., Fitzallen, N., English, L., & Wright, S. (2020b). Introducing statistical variation in year 3 in a STEM context: Manufacturing licorice. *International Journal of Mathematical Education in Science and Technology*, *51*(3), 354–387. https//doi.org/10.1080/0020739X.2018.1562117

Watson, J., Fitzallen, N., Fielding-Wells, J., & Madden, S. (2018). The practice of statistics. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 105–137). Springer. https://doi.org/10.1007/978-3-319-66195-7_4

Watson, J. M., & Kelly, B. A. (2002a). Can grade 3 students learn about variation? In B. Phillips (Ed.), *Developing a statistically literate society*. *Proceedings of the Sixth International Conference on Teaching Statistics,* Cape Town. International Statistical Institute. https://iase-web.org/documents/papers/icots6/2a1_wats.pdf?1402524960

Watson, J. M., & Kelly, B. A. (2002b). Grade 5 students' appreciation of variation. In A. Cockburn & E. Nardi (Eds), *Proceedings of the 26th Conference of the International Group for the Psychology of Mathematics Education*, Norwich, July 21–26 (Vol. 4, pp. 385–392). University of East Anglia.

Watson, J. M., & Kelly, B. A. (2002c). Variation as part of chance and data in grades 7 and 9. In B. Barton, K.C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds.), *Mathematics education in the South Pacific. Proceedings of the 25th Annual Conference of the Mathematics Education Research Group of Australasia, Auckland* (Vol. 2, pp. 682–289). MERGA.

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

Watson, J. M., & Kelly, B. A. (2005). The winds are variable: Student intuitions about variation *School Science and Mathematics*, *105*(5), 252–269. https://doi.org/10.1111/j.1949-8594.2005.tb18165.x

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, *34*(1), 1–29. https://doi.org/10.1080/0020739021000018791

Watson, J.M., & Moritz, J.B. (1998). Longitudinal development of chance measurement. *Mathematics Education Research Journal*, *10*(2), 103–127.

Watson, J.M., & Moritz, J.B. (2001). Development of reasoning associated with pictographs: representing, interpreting, and predicting. *Educational Studies in Mathematics, 48*, 47–81. https://doi.org/10.1023/A:1015594414565

Wild, C. (2006). The concept of distribution. *Statistics Education Research Journal*, *5*(2), 10–26. http://iase-web.org/documents/SERJ/SERJ5(2)_Wild.pdf

JANE WATSON
University of Tasmania
Private Bag 66
HOBART TAS 7000
AUSTRALIA

APPENDIX

Student Workbook Questions – Catapults Part 1 (space for response removed)

Today we are going to answer these questions about catapults:

- *How far does the Ping Pong ball travel?*
- *How consistent are we at launching the Ping Pong ball?*

Collecting Data using a Catapult

A1. What data can we collect?

A2. How can we make sure it is a *fair test*?

A3. How can we record the data?

[After carrying out trials and drawing a graph] *Questions about your Representation*

A4. How consistent were your data?

A5. What examples of variation are there in the data?

A6. Why might there be differences in the data?

Making Comparisons after creating a plot for the group's data

Now have a look at the graphs made by the other members of your group.

A7. Does your graph look the same as the graphs made by the other members of your group? Why/why not?

A8. Is one graph better than another at showing the variation in the data? Which one? Why/why not?

A9. Imagine you are going to launch your catapult one more time.

a. Predict how far it would go.

b. Describe how you used your graph to make your prediction.

Analysing Your Data Using the TinkerPlots Graph

A10. What do the data in the *TinkerPlots* graph tell you about your catapult trials?

A11. Are your data the same as the data in *TinkerPlots*? How can you tell?

A12. Describe any differences between the two representations.

A13. Is one representation easier to read and analyse? Why?

Conclusions about the Class Data

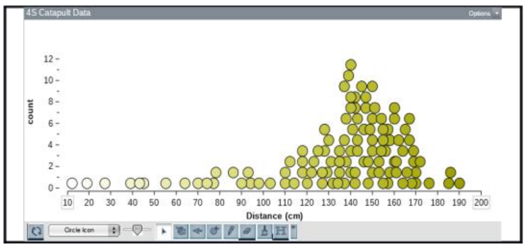
A14. Overall, how far did the Ping Pong ball travel?

A15. How consistent were we at launching the Ping Pong ball?

A16. What suggestions would you make to improve the consistency of the launches?

Student Workbook Questions – Catapults Part 2 (space for response removed)

What we did last time …



Using this *TinkerPlots* plot, tell the story about the data we collected as a class from the Catapults activity.

B1. What is the range of the data? __________ cm to ___________ cm

B2. How far did the ball typically travel?

B3. Where do you see *variation* in the plot? Give an example.

B4. Describe from the graph where the data are *consistent*.

Today we are going to collect data to answer the research question: How can you tell if the change you make to the catapult improves its performance? To make this a fair test we need to collect some data from the catapults before we make any adjustments. After adjustments are made, we will collect more data.

About Your First Data Trials

Look at the data in your table.

B5. Describe the *variation* you see in the data you just collected.

How *consistent* were you?

B6. How far did the ball typically travel?

Adjusting Your Catapult

To improve the catapult, each group will need to adjust the catapult by twisting the pegs to make the string tighter.

B7. How many turns of the pegs will each group make?

B8. From the data you collected for the *first trials*, predict how far you think the Ping Pong ball will travel *after you adjust* your catapult.

B9. How much further is it than your first trials?

About Your Second Data Trials

Look at the data in your tables.

B10. Did twisting the string improve the performance of the catapult?

□ Yes □ No

B11. How can you tell from the data you collected?

Using TinkerPlots to Analyse the Data from your Trials

B12. What is the *range* of the data in your plot from the two trials?

First Data Trials _________ cm to ___________ cm

Second Data Trials _________ cm to ___________ cm

B13. Describe the *variation* in your data.

B14. How far did the ball typically travel for the two trials?

First Data Trials ___________________________

Second Data Trials ___________________________

Use the Class Data to answer our Research Question: Have the catapults been improved?

B15. Using the graph of the *class data*, did changing the number of turns of the string improve the performance of the catapults?

□ Yes □ No

B16. How can you tell from the plot if the catapults have been improved or not?