A MODELING APPROACH TO THE DEVELOPMENT OF STUDENTS' INFORMAL INFERENTIAL REASONING

HELEN M. DOERR
Syracuse University, USA
hmdoerr@syr.edu

ROBERT DELMAS
University of Minnesota, USA
delma001@umn.edu

KATIE MAKAR
The University of Queensland, Australia
k.makar@uq.edu.au

Teaching from an informal statistical inference perspective can address the challenge of teaching statistics in a coherent way. We argue that activities that promote model-based reasoning address two additional challenges: providing a coherent sequence of topics and promoting the application of knowledge to novel situations. We take a models and modeling perspective as a framework for designing and implementing an instructional sequence of model development tasks focused on developing primary students' generalized models for drawing informal inferences when comparing two sets of data. This study was conducted with 26 Year 5 students (ages 10-11). Our study provides empirical evidence for how a modeling perspective can bring together lines of research that hold potential for the teaching and learning of inferential reasoning.

Keywords: *Statistics education research; Model-based reasoning; Informal statistical inference; Model development sequences; Model eliciting activities*

1. FROM INFORMAL INFERENCE TO MODEL-BASED REASONING

Research over the past decade has argued that curricula based on informal statistical inference have the potential to provide a more holistic and coherent approach to teaching and learning statistics (Bakker & Derry, 2011). There are several reasons for this. First, predictions are an everyday activity, allowing students to draw on their personal experiences to make meaningful inferences (Makar, 2013; Zieffler, Garfield, delMas, & Reading, 2008). Second, as Watson (2006) argued, creating an inference from data requires students to draw on the big ideas in statistics. This implies that informal statistical inference has great potential for generating opportunities to learn foundational statistical concepts (e.g., Garfield, Le, Zieffler, & Ben-Zvi, 2015; Makar, 2014). Third, informal inferential reasoning places the focus on the process (statistical reasoning) rather than only the product (conclusion). Statistical reasoning is what allows statistical concepts to be applied and used effectively in data-based arguments (Abelson, 1995; Ben-Zvi, 2006; Garfield, 2002). It is well documented that statistical reasoning is often missing from the teaching and learning of statistics (Ben-Zvi & Garfield, 2004). As Gigerenzer (1993) argued, "Statistical reasoning is an art … When it is mechanized, as with the institutionalized hybrid logic, it

becomes ritual, not reasoning" (p. 333). Together, these three ideas—making personal meaning, integrating core statistical concepts, and focusing on reasoning—explain why informal statistical inference provides a powerful tool for creating a holistic and coherent approach to learning statistics.

We see informal statistical inference as being characterized by: (1) a claim beyond the data; (2) the use of data as evidence for the claim; (3) an expression of uncertainty (Makar & Rubin, 2009). Researchers have argued that informal inferential reasoning, the reasoning that leads to informal statistical inference, is promoted by an inquiry-based environment that engages students with statistical tools, promotes norms of collaboration, and involves solving intriguing tasks that stimulate reasoning (Makar, Bakker, & Ben-Zvi, 2011; Paparistodemou & Meletiou-Mavrotheris, 2008; Zieffler et al., 2008).

The opportunity to learn statistics by focusing on inferential reasoning is an important shift in direction for the field in light of Bakker and Derry's (2011) three key challenges in statistics education: (1) avoid *inert knowledge*, where students can reproduce what they have learned but not apply their knowledge to novel statistical problems; (2) avoid *atomistic approaches* and foster *coherence* in curriculum from students' perspectives; and (3) *sequence* topics in new ways to improve coherence from students' perspectives. Although informal inferential reasoning may have moved the field forward with respect to Bakker and Derry's second challenge of *coherence*, theoretical work is still needed to address the first (*inert knowledge; applying knowledge to novel statistical problems*) and third challenges (*sequence topics for coherence*). We believe that modeling can play an important role in extending the informal statistical inference framework to address these remaining two challenges by focusing on important elements that are left ambiguous in research on informal statistical inference—purposeful collection of data, development of useful representations, and applying disciplinary structures to new problems. Our study illustrates how a perspective of model-based reasoning in statistics can bring these elements to the foreground more explicitly in the teaching and learning of statistics.

In this article, we present the results of a study conducted by the authors within an inquiry-based primary classroom in Australia. The study adopted Lesh, Cramer, Doerr, Post, and Zawojewski's (2003) ideas of teaching through a sequence of modeling activities to develop primary students' informal inferential reasoning about comparing two distributions, an important structural model for solving a broad class of problems in statistics. Through this study, we argue for the importance of adopting a model-based approach to instruction to meet the challenges posed by Bakker and Derry (2011). The central research question we address in this article is: *How did the model development sequence support students in developing a generalizable model for comparing two groups of observed data?*

1.1. PERSPECTIVES ON REPRESENTATIONS

Statistical modeling has typically been a topic addressed in advanced courses, although its use from kindergarten through introductory statistics has been an emerging development (e.g., delMas, Garfield & Zieffler, 2014; English, 2012; Fielding-Wells & Makar, 2015; Konold & Harradine, 2014; Lehrer, Jones & Kim, 2014; Lehrer, Kim & Jones, 2011; Makar, 2016; Peters, 2011). Our literature review focused on research relevant to the context of our study: developing the model-based reasoning of primary school children. We found resonance with informal inferential reasoning and informal statistical inference in Lehrer et al.'s (2011) characterization of data modeling: "Data modeling integrates inquiry, the generation of data, chance and inference" (p. 725). Our intent, therefore, was to seek literature that aimed for a broad interpretation of modeling that would capture

informal statistical inference, the generation of data through inquiry-based activities involving uncertainty, and the representations of data to support analysis and conclusions.

Representations do not simply "represent" but also carry with them mathematical, cognitive and cultural activity (Font, Godino & D'Amore, 2007). Conventional representations often evolve out of a long history from repeated sequences of personal localized objects applied across different problems and contexts, until some characteristics become stable and then invariant as they are institutionalized within the disciplines of mathematics and statistics. The student is not usually privileged to experience this process and, therefore, does not have the opportunity to develop a personal meaning for the representation. Nonetheless, the student is expected to carry disciplinary meanings into an application of a representation to a local problem.

Any personal meaning for representations arises "through a considerable amount of situated, lived work" (Roth & McGinn, 1998, p. 40). Referring to representations as inscriptions, making meaning of an inscription depends on "the degree to which individuals participate in purposive, authentic, inscription-related activities" (Roth & McGinn, p. 37). An inscription does not have its own meaning; its meaning arises from the context of its use and within the context of other inscriptions and language. Representations that emerge from activity may be "personal objects," or they can be "institutional objects" if the practices that they represent are shared, or a hybrid of the two (Font et al., 2007). Nemirovsky and Tierney (2001) explained that when children add names and distinctive marks (with a key) to a graph or list data sequentially rather than using spacing, they demonstrate their need to hybridize between disciplinary aspects of a representation (number line with points ordered left to right) and personal ones (recognizable aspects or omissions that are not needed such as empty spaces where there are no data). Their hybrids allow children to preserve information that is important to them. We see representations as emerging from activity in mathematics and statistics, which once negotiated can be abstracted to be useful across a number of contexts.

1.2. PERSPECTIVES ON MODELS AND MODELING

Models have a similar evolution to representations in emerging out of human activity. Hestenes (2010) defined a model as a representation of structure in a given system, where a system consists of a set of related objects and the structure identifies the relationships among the objects. A model creates a sharable representation that provides a common language and referent for discussion. Working in scientific modeling, Hestenes argued that curriculum should be organized around a set of conceptual models as the basic units of scientific knowledge, with a small set of basic models as the core content. He further argued that theory should be taught as general principles for constructing models and that students should learn a modeling approach to scientific inquiry, which consists of developing proficiency with conceptual modeling tools, qualitative reasoning with models, methods for quantitative measurement, and facility with comparing models to data. The role and responsibility of the teacher, therefore, is to frame classroom activity in terms of models and modeling. An end goal of modeling activities is to build shared meanings of phenomena under investigation (i.e., the system that is modeled) through the comparison and evaluation of various models. Models involve more than just representations. To Hestenes, a model captures three important elements: (1) simplicity, to "minimize superfluous or empty relations" (p. 26) by focusing on aspects of the phenomenon that the modeler cared about and diminishing the rest; (2) generalizability, by capturing the structure of the phenomenon; and (3) applicability, seeing a model as providing a purposeful way of making sense of the world through applications to authentic problems.

Similar to Hestenes' (2010) perspective, Lehrer and Schauble (2010) referred to model-based reasoning as a type of explanation that is characteristic of science. From their perspective, models represent "analogies in which objects and relations in one system, the model system, are used as stand-ins to represent, predict and elaborate those in the natural world" (p. 9). Lehrer and Schauble stated:

Model-based reasoning entails deliberately turning attention away from the object of study to construct a representation that stands in for that phenomenon by encapsulating and enhancing its theoretically important objects and relations. Instead of directly studying the world, one studies the model – the simplified, stripped-down analog. (p. 13)

They contended that there is a wide range of representational forms, which comprise a language of expression for model-based reasoning, and include drawings, diagrams, maps, physical replicas, mathematical functions and simulations. They consider the invention and use of this "representational language" to be a critical precursor to students' development of model-based reasoning. Lehrer and Schauble argued for experiences that help students develop a broad and powerful vocabulary for mathematizing situations. Students in model-based reasoning activities are expected to build models (e.g., data representations in statistical investigations) that are convincing to other investigators (e.g., fellow students) and determine how much trust should be given to conclusions derived from the model.

According to Lehrer and Schauble (2010), younger students do not naturally accept the representational validity of a model. For younger students, to represent is to copy, seeing is believing, and it is not obvious why a model needs to be constructed for something that can be observed directly. Because all representations and models involve trade-offs, students need experience with evaluating the pros and cons of various representations against the purpose of an investigation or task. This can lead to an understanding that theory can be regarded as hypothetical and needs to be evaluated against evidence. To this end, Lehrer and Schauble promoted the use of tasks that can be addressed with a variety of representations or representational forms.

When students work with statistical models, the aim of a modeler is not to describe the data in front of them but rather to use the data they have to draw a conclusion about the context represented by the model (Pratt, Johnston-Wilder, Ainley, & Mason, 2008). Modeling (or model-based reasoning) includes an evaluation of both the fit and the misfit of a model, which can involve statistical ideas of sampling, variability, and uncertainty. Model evaluation drives model revision, which generally increases the explanatory power of a model and the scope of application as the model is fine-tuned to account for new data (i.e., evidence in scientific inquiry). From an instructional perspective, teachers need to identify modeling contexts that afford or promote the evaluation of models.

Barbosa (2006) offered some additional perspectives on modeling activities with students. From Barbosa's perspective, modeling must contain elements of purpose and authentic connection to context. He defined the boundaries of a modeling activity as having two main features:

the activity has to be a problem (not an exercise) for the students;

the activity has to be extracted from the everyday or other sciences that are not pure mathematics. (p. 294)

Barbosa's work speaks to the importance of embedding the context of the generation and application of the model in the everyday experience of the student so that modeling does not become simply an "exercise". He suggests that in studying students' modeling practices, "Rather than measuring how closely students approximate normative descriptions of modeling, their practices can be viewed as fertile ground for formulating theory about modeling in the schools, the conditions of its production, and the possibilities for constituting socio-critical practices" (p. 296-7). For Barbosa, the interest is not in

comparing student ideas to expert practices, but in studying the students' development from their own experiences and within their school context. Barbosa's work highlights that the *modeler* is part of the modeling process. This requires modeling activities that value the problems faced by the modeler, the contexts being modeled, and the modelers themselves.

1.3. MODEL DEVELOPMENT SEQUENCES

A models and modeling perspective (Doerr & English, 2003; Lesh & Doerr, 2003) presents a framework for designing instruction that incorporates the modeling theory, model-based reasoning, and modeling activity perspectives of Hestenes (2010), Lehrer and Schauble (2010), and Barbosa (2006). From a models and modeling perspective, a model is defined as a system consisting of elements, relationships, rules and operations that can be used to make sense of, explain, predict, or describe some other system (Doerr & English, 2003; Lesh & Doerr, 2003). An underlying assumption of this perspective is that learning occurs through the process of developing an adequate and productive model that can be used and re-used in a range of contexts. Similar to Lehrer and Schauble, the invention and use of representational forms (such as drawings, diagrams, maps, physical replicas, mathematical functions and simulations) are critical steps that learners need to take to develop model-based reasoning about phenomena.

One category of modeling tasks that has been studied over a range of content areas is known as model eliciting activities (MEAs). These activities are designed to elicit students' initial conceptions and representations about realistic and meaningful problem situations. Often students' initial ideas are not very useful or sophisticated. However, as students discuss their approaches with other students or share them with the whole class, they have the opportunity to evaluate and reject initial conceptions and to revise them in ways that are increasingly productive and accompanied by useful representations. MEAs encourage groups of students to engage in an iterative process where they express, test, and refine their ways of thinking about meaningful problem situations. MEAs are designed to elicit a generalizable model that reveals the underlying structure of the problem situation.

Six principles for guiding the design of an MEA developed by Lesh and colleagues have been widely used to develop MEAs at all grade levels K-16 and in many content areas (Lesh, Hoover, Hole, Kelly, & Post, 2000). Each principle is met by designing an activity to address a set of questions as follows:

(1) *The Reality Principle:* Could this problem really happen in a real-life situation? Will students be encouraged to make sense of the situation based on their personal knowledge and experiences?

(2) *The Model Construction Principle:* Does the task ensure that students will recognize the need for a model to be constructed, modified, extended or refined? Does the task involve describing, explaining, manipulating, predicting or controlling some other system? Is attention focused on underlying patterns and relationships, rather than on surface features?

(3) *The Model Documentation Principle:* Will the responses that students generate explicitly reveal how they are thinking about the situation? What kinds of mathematical objects, relationships, operations, patterns and regularities are they thinking about?

(4) *The Self-Evaluation Principle:* Are the criteria clear to students for assessing the usefulness of alternative responses? Will students be able to judge for themselves when their responses are good enough? For what purposes are the results needed? By whom? When?

(5) *The Model Generalization Principle:* Can the model that is elicited be applied to a broader range of situations? Students should be challenged to produce re-usable, shareable, and modifiable models.
(6) *The Simple Prototype Principle:* Is the situation as simple as possible, while still creating the need for a significant model? Will the solution provide a useful prototype for interpreting a variety of other structurally similar situations?

However, a single MEA is not enough for students to develop fully a generalized model that can be used in a range of contexts (Ärlebäck, Doerr, & O'Neil, 2013; Doerr & English, 2003). To achieve this goal, students need to engage in a sequence of model development activities (see Figure 1). Such sequences are structurally related tasks, beginning with an MEA and followed by one or more model *exploration* activities (MXAs) and model *application* activities (MAAs). These sequences provide a way of organizing instruction on a central concept such as inferential reasoning or variation and distribution. The MEA elicits students' initial models. The MXAs engage the students in thinking *about* the models that were elicited. The MXAs focus on the underlying structures of the models and on the strengths of various representations and ways of using them productively. After thinking about their models, MAAs engage students in thinking *with* their models by applying them to new contexts. As Lehrer and Schauble (2010) pointed out, "A good representational solution for one kind of problem may not work well for another" (p. 16). MAAs engage students in making adaptations to their models, extending or modifying previously explored representations, and refining language for describing and explaining phenomena. Each component of a model development sequence (Figure 1) engages students in multiple cycles of descriptions, interpretations, conjectures, and explanations that are revised and refined while working with other students. As illustrated in Figure 1, the initial MEA can be followed by an MXA or an MAA, which can then be followed by any order of additional MXAs and MAAs that promote the development of a targeted conceptual system.

Model Development Sequences

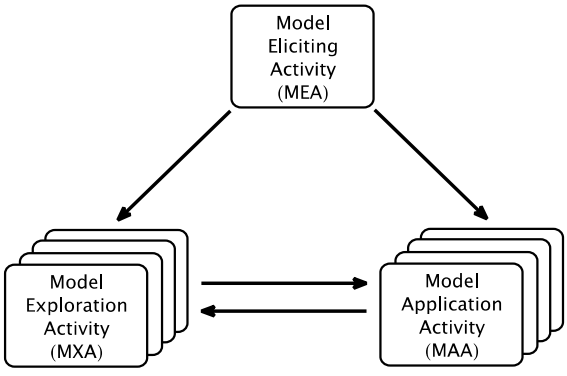


Figure 1. The general structure of a model development sequence

2. METHODOLOGY

2.1. DESIGN AND PARTICIPANTS

In this collaborative research study, we draw on a view of educational design research that focuses simultaneously on the design of a product (in this case the design of an instructional sequence focused on developing students' generalized models for drawing

informal inferences when comparing two sets of data) and on the generation of theory or principles about the teaching and learning of model-based inferential reasoning. A design experiment (e.g., Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003; Gorard, Roberts, & Taylor, 2004; Hovan & Palalas, 2011) approach was used, developing the activities through an iterative process of initial task design based on theory, field testing, updating theory, redesigning the task, and repeating the process. As Kelly, Baek, Lesh, and Bannan-Ritland (2008) observed: "design researchers may learn not only how to improve an innovation, but also how to conduct just-in-time theory generation and testing within the context of design processes and in the service of the learning and teaching of content" (p. 5). We are explicitly locating our work in the messy context of a classroom, with all of its constraints, as the research team itself worked to understand a modeling approach to informal inferential reasoning and the application of task design principles to the innovation.

Participants were 26 Year 5 students (18 males and 8 females, ages 10-11) in an inquiry-based classroom from a primary school in a suburb of Brisbane, Australia. One member of the research team (Makar), who knew the classroom teacher and her students, taught the instructional sequence. The research team met with the classroom teacher, who also took an active role in the teaching and contributed much to the work of the team. Similar to other work in the tradition of design-based research, the research team met after each lesson, shared observations about students' emerging ideas, identified modifications to be made in the instructional activities, made decisions about which student representations to draw on in class discussions, and shared ideas about questions to ask the students and general teaching strategies. In this way, cycles of testing and revision occurred as we implemented the innovation (the instructional sequence) and theory building occurred as we sought to understand what a modeling approach brings to the development of students' informal inferential reasoning.

2.2. THE MODEL DEVELOPMENT SEQUENCE

The modeling activities were based on the paper helicopter experiments introduced by Box (1992). The research team reviewed articles on paper helicopter experiments (e.g., Ainley, Pratt, & Nardi, 2001; Annis, 2005; Erhardt, 2007; Siorek & Hafta, 1998). The overall learning goal of the sequence of activities (MEA, MXA, MAA) was to have students develop a model for comparing groups; the purpose of each modeling activity is shown in Table 1.

Table 1. The three types of modeling activities

Modeling Activity	Abbreviation	Purpose
Model Eliciting	MEA	Elicit initial conceptions and representations
Model Exploration	MXA	Engage students in thinking about the models
Model Application	MAA	Students adapt, modify, or extend models

Field notes taken during each lesson and student-generated artifacts (e.g., dot plots, written predictions, written conclusions) were analyzed and discussed by the researchers to inform modifications to the next lesson prior to implementation. We briefly describe each of the activities that comprise the model development sequence.

The model eliciting activity (MEA) The initial MEA engaged the students in investigating the question: *How long does a paper helicopter stay in the air?* The intent of the investigation was to have students aggregate, describe, and explain the data from multiple groups of students collecting data about their paper helicopters. Rather than specify a particular procedure for collecting and representing the data, we planned to have

the students create identical helicopters (see Figures 2 and 3) and to make available a range of tools, including meter sticks and stop watches, for collecting the data. No instructions were given on how to "launch" the helicopters or on how to measure flight durations as we wanted to begin with students' initial conceptions and to support them in developing a need for a process to generate, collect, and represent data in order to respond to the question under investigation.

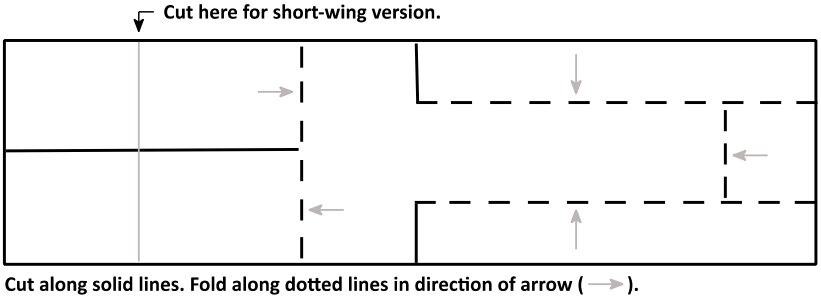


Figure 2. Template students used to create a paper helicopter

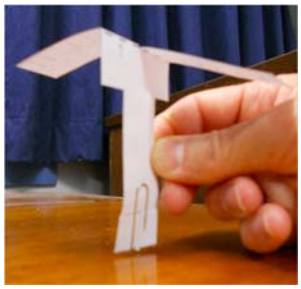


Figure 3. An assembled paper helicopter

The model exploration activity (MXA) Dot plot representations were new to this group of students. Building on the outcome of the MEA, the MXA focused on the creation and interpretation of meaningful dot plots as a first type of representation that would allow the students to address the inquiry question. Following from a discussion of aggregated data from the MEA, the students were given the following task: *Collect enough data from two different heights to show evidence with dot plots that height matters.* The intent of this MXA was on guiding the students to explore the usefulness of dot plots as a means to compare two groups, thus drawing on Hestenes' (2010) ideas of model (e.g., a representational structure that identifies relationships among a set of objects involving quantitative measures and to which qualitative reasoning can be applied). Students worked in small groups of three to four to collect data and create dot plot representations, which were publicly displayed on poster paper. The researcher, with the aid of the classroom teacher, led a whole class discussion where students (a) interpreted the data and (b) evaluated the representations by comparing and identifying features that facilitated comparison of flight durations for the two heights. This discussion was designed to promote Hestenes' and Lesh et al.'s (2003) modeling activity goals of building shared meanings and evaluating models.

The model application activity (MAA) The MAA had students work in pairs to apply what they learned from the MEA and MXA to design and carry out a method for collecting evidence of a difference in flight durations between two types of helicopters – helicopters with long wings and with short wings. They were told the manufacturer of the paper helicopters needed an answer to the following question: *How big of a difference is there between the flight duration of short-wing and long-wing paper helicopters?* Their system (or model) for answering this question had to be described in such a way that the manufacturer could carry out their plan for studying this question. In addition to providing an estimate of the difference in flight durations, the student groups had to provide a statement of "how sure" they were about the size of the difference. Students were encouraged to include graphs or pictures that would help the manufacturer understand their estimate. The students worked in pairs with two helicopters, one long-wing and one short-wing, and collected data to draw conclusions and support their conclusions with evidence.

2.3. DATA COLLECTION AND ANALYSIS

Data collected during the activities described above included video recordings of each session, field notes, students' recorded predictions and conclusions in the MEA, dot plots produced during the MXA, and the comparative dot plots made during the MAA. These artifacts were analyzed to address our central research question: How did the model development sequence support students in developing a generalizable model for comparing two groups of observed data?

The artifacts and videos were analyzed, first separately and then increasingly in an intertwined process. The video analysis followed a process similar to the seven-stage model described by Powell, Francisco, and Maher (2003): attentive viewing, description, identifying critical events, transcribing, coding, construction of a storyline, and composing narrative. Although the analysis is described here as an ordered process, in practice it was non-linear and overlapping. Some stages were begun but then left incomplete and returned to later to fill in gaps or temporarily skipped if needed. The drafted narrative, as explained by Powell et al., was developed from the beginning alongside the other stages of analysis to assist with meaning-making, and clarifying or modifying key ideas as insights were gained from analysis and discussion among the authors.

The researchers each watched the three lessons multiple times to re-familiarize themselves with the general content of the lessons. Although all three authors were present during the lessons, viewing of the video assisted with clarifying details from a common source. A brief video log was created including timestamps, screen shots, and short descriptions of lesson segments to facilitate the location of critical events. The major learning objectives were identified for each of the three lessons. The beginning and end point at which a change or shift occurred in classroom discourse that moved the class in the direction of a learning objective was called a milestone. Five milestones were identified across the three lessons (see Table 2).

Table 2. Major milestones across the model development sequence

Milestone	Description	Activity
1	Attending to variation	MEA
2	Collectively representing evidence	MXA
3	Does drop height matter?	MXA
4	Identifying useful characteristics of representations for comparing groups	MXA
5	Applying the students' model for comparing groups	MAA

The field notes and video logs were used to identify segments of the video recording of a lesson where a milestone occurred. Narrative accounts consisting of transcriptions and still video frames were created to produce a storyboard for each milestone. The narrative transcripts were used to identify how students interacted with the paper helicopters, and the nature of interactions between the teacher/researcher and the students. After each lesson, the student-generated artifacts (dot plots, written predictions, written conclusions) were analyzed to identify characteristics of students' representations and thinking about comparing groups across the model development sequence. The field notes, narrative transcripts, and artifact analyses were used to provide evidence of the development of the students' understandings relative to each milestone.

The main artifacts from the MEA were the students' initial predictions, their estimated flight times and statements of conclusions after experience with measuring the flight times of the paper helicopters, and a whole class dot plot of measured flight times. Predictions were coded for whether or not they represented an interval estimate. Conclusions were coded for whether or not estimates were stated as a type of average (using the words "average," "about," or "around"), a statement of clustering (stating where "most" or "more" of the flight times were located), if variability in the flight times was noted, and if a statement indicating that the flight times depended on the drop height was included.

The main artifacts from the MXA were conclusions written by each student after a discussion of the whole class dot plot from the MEA and dot plots produced by each small group of students after collecting flight times from two different drop heights. Students' individual conclusions were coded for whether or not they: explicitly stated that flight time depends on drop height, provided a range or interval estimate for flight time duration, based their estimated flight times on where values clustered, and stated that the original inquiry question was not specific or that there was not a single answer to the inquiry question. The group dot plots were coded for the number of points on each dot plot, if two number lines were used (one for each drop height), and if number lines were divided into equal intervals or if flight times were placed in rank order along a number line. If two number lines were used, dot plot representations were coded for whether or not the number lines were aligned vertically and if the same scale was used on each number line (e.g., both went from 0 to 2 seconds).

The main artifacts collected from the MAA were the dot plot representations and conclusion statements produced by each small group to compare the flight times of short-wing and long-wing paper helicopters. The same characteristics coded for in the MXA dot plots were coded for in the MAA dot plots. In addition, dot plots were coded for whether or not an "average" was reported for each type of helicopter, the estimated difference in flight times, and the degree of overlap between the distributions for the two types of helicopter (complete or almost complete, some, or none). The conclusion statements were coded for whether or not the degree of overlap was acknowledged and whether or not a group stated that there was a difference between the flight times of short-wing and long-wing helicopters.

3. RESULTS

We draw on the empirical data from this study to address our central research question: How did the model development sequence support students in developing a generalizable model for comparing two groups of observed data? We report our findings in terms of five milestones that occurred over the three days of the instructional sequence, beginning with our analysis of the MEA. Students' conclusions are reported exactly as they were written,

which includes their misspellings of words. Each student's real name has been replaced by a pseudonym.

3.1. MILESTONE 1: ATTENDING TO VARIATION

We wanted to begin our research on the development of students' models for comparison from their initial strategies for representing and comparing data. We wanted to create a need for collecting, representing, and interpreting data that would respond to the inquiry question about flight times for helicopters and ultimately to support students in using data representations to make meaningful comparisons. The MEA engaged the students in playing with paper helicopters to investigate the inquiry question posed by the teacher: How long does a paper helicopter stay in the air? Twenty-three of the 26 students were in attendance during the MEA. Before actually playing with the helicopters, the students were shown a typical paper helicopter and asked to predict how long it would stay in the air (see Figure 3). There was a large range in the first prediction given by each student, from a low of two seconds to a high of one minute (see Table 3). Only three students provided an interval for their predictions (2-5 seconds; 3-5 seconds; 30-45 seconds), thus the majority of students provided a point, or single value, prediction. This initial prediction was intended to focus students' attention on the inquiry question: the length of time a paper helicopter would stay in the air.

Table 3. First predictions for how long a paper helicopter stays in the air

	Estimate Type		Total
	Single Value	Interval	
Prediction	N (%)	N (%)	N (%)
Value(s) between 2 and 6 seconds	7 (30%)	2 (9%)	9 (39%)
Value(s) between 10 and 15 seconds	7 (30%)	0 (0%)	7 (30%)
Value(s) between 30 and 60 seconds	6 (26%)	1 (4%)	7 (30%)
Total	**20 (87%)**	**3 (13%)**	**23 (100%)**

Following their written predictions, students created and launched their paper helicopters (see Figure 2) to collect evidence to answer the inquiry question. After working in small groups to make sense of the flight times of the paper helicopters, students were again asked to write down their prediction about how long a paper helicopter would stay in the air. These estimates, based on their observed data, not surprisingly were less variable than their first predictions, ranging from 1.24 to 3 seconds. The most common response was a point estimate of 2 seconds given by 12 students (52%). Three other students (13%) gave intervals that had a midpoint of 2 seconds (one student wrote 0.20 – 3.8 seconds, and the other two each estimated 1 to 3 seconds). Of the remaining students, six (26%) gave a point estimate between 1.24 and 1.8 seconds, and two others gave point estimates of 2.44 and 3 seconds, respectively. In contrast to the first predictions, seven (30%) of the students provided an interval of values for their estimate. After playing with the paper helicopters, there was less variation in estimated flight times relative to the predicted flight times, and more students expressed their second predictions as intervals.

Each student also wrote a conclusion about the investigation of the flight times. Although the majority of students still provided point estimates for the flight times, nine (39%) of the conclusion statements qualified their estimates with words and phrases such as "about," "around," "mostly," "most people," "more times," or "most of the time." Examples of statements that identified a qualified estimate are:

Ankita: The results show that the average amount of time the paper helicopter stayed in the air for was about 2 seconds. The way we found out this data is we had a results page and measured the height and how many seconds it stayed in the air and recorded it.

Noah: Our groups result was around 1.25. All three of us stood as high as we could then dropped it while another member used a stop watch to time.

Anna: Mostly 2 seconds. The hellicopter stays in the air for 2 seconds most of the time but depends on the height.

Jordan: 2.0 seconds because there were more times in the 2.00 sec mark

An additional 11 students (48%) described or stated a range of values, or stated that the values were not all the same. Examples of these types of statements are:

Simon: The paper helicopter was in different time all round, because we dropped it at different heights so it had different times

Riley: Our group found out that the paper helicopter stayed in the air for 3 seconds, the most and 2 seconds the least.

Vijay: Our group found out that our paper helicopter stayed in the air for 1-3 seconds when [TEACHER] stood on a chair and dropped it at the ceiling height.

Wei: Between 0.20 seconds and 3.80 seconds, the helicopter had stayed in the air.

Thus, a majority of the students expressed recognition of the variability in the flight times even though a majority of the estimates were reported as point estimates. At the same time as students recognized the variability in flight times, new ideas about sources of variation were elicited by the task. Our viewing of the video recording of the students' play with the helicopters revealed a variety of ways in which the paper helicopters were launched. For example, some groups threw the helicopter up into the air resulting in varying heights from which the helicopter started to fall. Others dropped the helicopter from varying heights, with some of these groups using measuring sticks to determine and record the drop height. In their written conclusions about their investigation of the flight times, 11 of the students (48%) made a statement indicating a relationship between the height at which the paper helicopter started and the flight time. Below are some examples of statements that indicated recognition of the relationship between these two quantities:

Matt: I think that the higher the helicopter gos the longer it flys in the air. If the helicopter is thrown in the air it took about 3 seconds to reach the ground.

Martin: The paper helicopter varried in time because there needs to be a specific height to drop from and there was no particular height. The higher the helicopter was dropped from the more time it took to get to the ground.

Beth: Our results prove the time a paper helicopter stays in the air is determined by the height that we through [threw] it from and the way the helicopter is throughn [thrown].

The variation that occurred in how the students launched the helicopters into flight made visible to the students that the height from which the helicopter is dropped matters in terms of the flight time. Thus, at the end of this first task, we found that the students were beginning to shift from point estimates of flight times to interval estimates and to a recognition that the drop height is now a variable of considerable interest for answering the model eliciting question about how long a paper helicopter stays in the air.

3.2. MILESTONE 2: COLLECTIVELY REPRESENTING EVIDENCE

The students' work on the MEA created the need for the students to look collectively at their data on flight times for helicopters. To meet this need, the teacher introduced a common number line on which students could record their data, as shown in Figure 4.

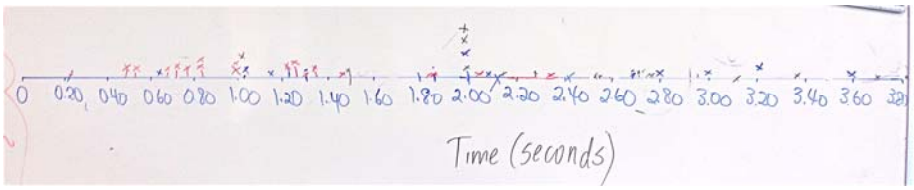


Figure 4. A number line for sharing students' data on flight times

This in turn led to a whole class discussion that focused on the interpretation of the representation and an evaluation of the representation's usefulness in supporting the drawing of conclusions from evidence. The students' interpretations and evaluations of this representation formed the core of the MXA, as described earlier. All of the 26 students were in attendance during the MXA. The teacher drew the students' attention to the variability in the flight times and asked if the class could give an answer to the inquiry question: "How long does a paper helicopter stay in the air?" There was a mixture of "yes" and "no" responses from the students. The students then paired up and discussed whether or not the dot plot allowed them to come to a conclusion. After about five minutes of discussion, the teacher directed the students to each write down a conclusion, emphasizing that students needed to explain why they came to their conclusions.

Our analysis of the students' written conclusion statements showed that 17 (65%) of the 26 students made a statement that was coded as "height matters" with respect to flight times. Examples of such statements are:

Lucy:	It depends on what height you drop it from because if it was a hiar height it could go for 1 second and if you drop it from the roof it could be 3 seconds.
Jack:	I think that it would be Beetween 1 sec and 3 secs But it Depends on how high you throw the helicopter.
Sarah:	If we all put it from the same hight we might get about the same answers.

These statements suggest that students recognized that different groups of students dropped the helicopters from different heights and that this variability in drop height contributed to the wide variation in the flight times that students recorded. Four of the students who made a "height matters" statement also indicated that the inquiry question did not specify a drop height and two other students noted that there was no single answer to the inquiry question. Thus, the students concluded that "drop height matters" and that the dot plot representation did not adequately capture the variability in drop height.

We also note that in their second set of conclusions, after discussing the whole-class dot plot, the students continued to provide both point (single value) estimates and interval estimates (see Table 4). Half of the students stated an interval, with three students essentially stating the flight times could be anywhere from 0 to infinity, while the other 10 students expressed intervals within the range of observed flight times. Of the nine students who provided single value point estimates, six made qualifying statements that their point estimate was based on where most of the values were on the dot plot or recognized that there was a range of values.

Table 4. Type of predictions made in students' second conclusions

Type of Predication	N (%)
Interval estimate	
Anywhere from 0 to infinity	3 (12%)
Interval within the range of observed flights	10 (38%)
Total	**13 (50%)**
Point estimate	
1 second	2 (8%)
2 seconds	6 (23%)
1.05 seconds and 2.34 seconds	1 (4%)
Total	**9 (35%)**
No estimate	
Because flight time depends on the height	3 (12%)
Statement not relevant to the inquiry question	1 (3%)
Total	**4 (15%)**
Total	**26 (100%)**

Four of the students did not provide an estimate. Three of these students stated that they could not give an estimate because the flight time depends on the height at which the helicopter was dropped:

Angus: I think there is no average because you can drop the helicopter form 1m and get 2 sec but you can drop off a building it could take 1 hour so it depends on the height you drop it from.

Riley: It depends on the hight of were your droping it from e.g. if you drop it from a 10m tree it might be 10 or 20 seconds.

Simon: My conclusion is that I can't answer because people got different times at different heights cause the question was not very specific.

The conclusion written by the fourth student did not address the inquiry question ("The height of the helicopter when you drop it"). Thus, a majority of the students indicated that they were aware of the variability in the flight times and that the variability was related to the height at which a helicopter was dropped or launched.

3.3. MILESTONE 3: DOES DROP HEIGHT MATTER?

The students' conclusion that drop height matters when answering the question posed in the MEA (How long does a paper helicopter stay in the air?) led the research team to design a new question for the teacher to pose in the MXA: "Drop height seems to matter. How can we tell this from the class data?" This question continued the MXA by pressing for the collection and representation of data that would provide evidence to answer this new inquiry question. The teacher recorded the students' conclusion on the board in the front of the classroom: "the helicopter will stay in the air longer if it is dropped from a higher height." She then posed a new task for the students that focused the students' attention on collecting evidence from two different heights and on creating a representation that would allow the students to draw conclusions based on their data or evidence: "Collect enough data from two different heights to show evidence with dot plots that height matters." As with the previous inquiry question, the students were not given specific heights to use, nor were they told how much data to collect, nor were they told how to represent their data on dot plots. Not surprisingly, this MXA led to more variation as both drop heights and the number of drops made when collecting evidence varied across student groups.

Our analysis of each group's dot plots showed noticeable variation in how the distributions of flight times were displayed. The seven groups did not use exactly the same two drop heights. Four of the groups used 1m and 2m as the respective drop heights. The other three groups used drop heights of 0.5m and 1m, 1.5m and 2m, and 1.44m and 2.67m, respectively. Additionally, the number of drops from each height varied considerably, with six of the groups making between 5 and 15 drops from each height, and one group that made only two drops at each height.

All of the dot plots were based on horizontally oriented number lines, similar to the number line that the teacher provided for the whole class dot plot of flight times shown in Figure 4. Five of the seven groups used two separate number lines, one for each of the two drop heights, and aligned the two number lines vertically. Among these five pairs of dot plots, only two of the groups used the same scale for both dot plots (e.g., the number line for both dot plots went from 0 to 2 seconds) with each number line drawn to approximately the same physical width, and with equal intervals as shown in Figure 5.

a.

b.

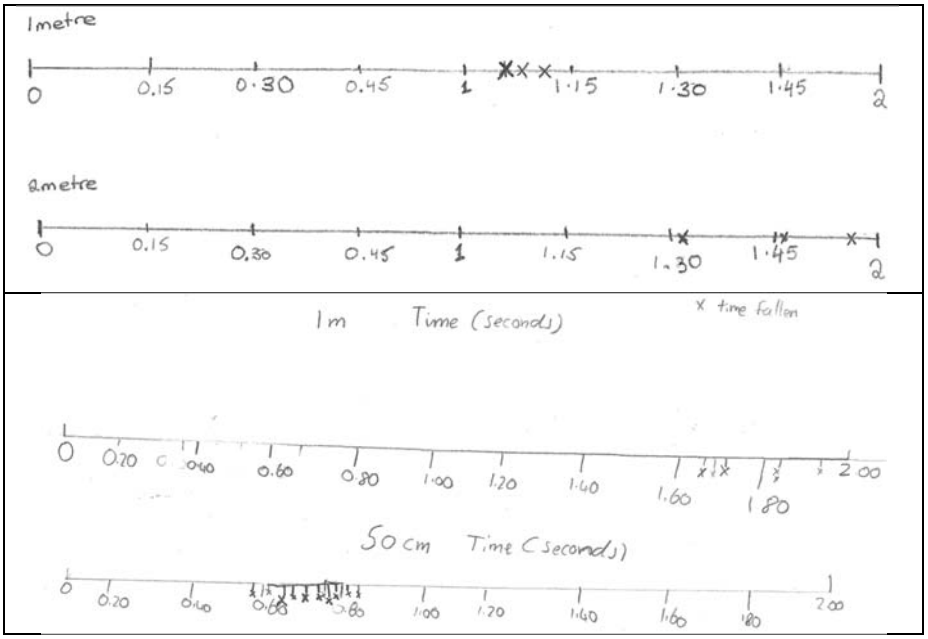


Figure 5. Dot plot pairs where with the same scale for each dot plot

Two other groups marked off equal intervals on the number lines. One of these groups used two vertically aligned number lines where the physical width of both number lines were the same but the intervals were not (from 0 to 2 seconds for a 1m height, but from 1 to 2 seconds for a 2m height; see Figure 6). The other group used a single number line that went from 1 second to 2.4 seconds marked off in 0.1-second increments (Figure 7). This group made two drops at each drop height, but only a single flight time was plotted for each drop height. During class discussion the students in this group said that the flight times for both drops at a particular height were very close to the plotted point.

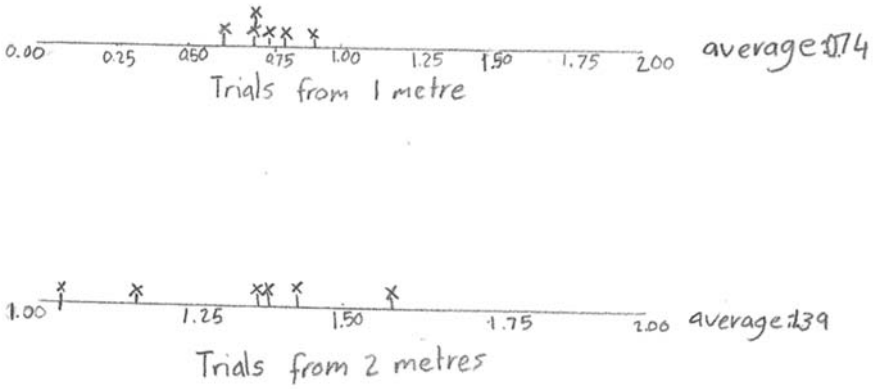
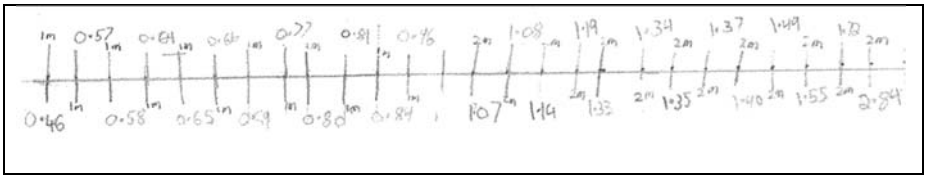


Figure 6. Dot plot pair with a different scale for each dot plot

Figure 7. Dot plot with a single number line

Of the remaining three groups, two groups placed all or some of the flight time values in rank order along their respective horizontal lines. One of these groups used only a single line to rank order 12 flight times at a 1m drop height and 13 flight times from a 2m drop height (see Figure 8a). All of the 1m flight times were less than all of the 2m flight times. A vertical line indicated each flight time with the drop height recorded at one end of the line (e.g., the top) and the flight time recorded at the other end (e.g., the bottom), alternating the positions of the labels from one vertical line to the next. The vertical lines were evenly spaced so that the distance between the vertical lines did not represent actual differences between flight times. The other group used a separate horizontal line for each drop height (Figure 8b). The two lines were of the same physical width but had different start and end values. Flight times were placed along each horizontal line in rank order, but they were not located relative to the end points.

a.



b.

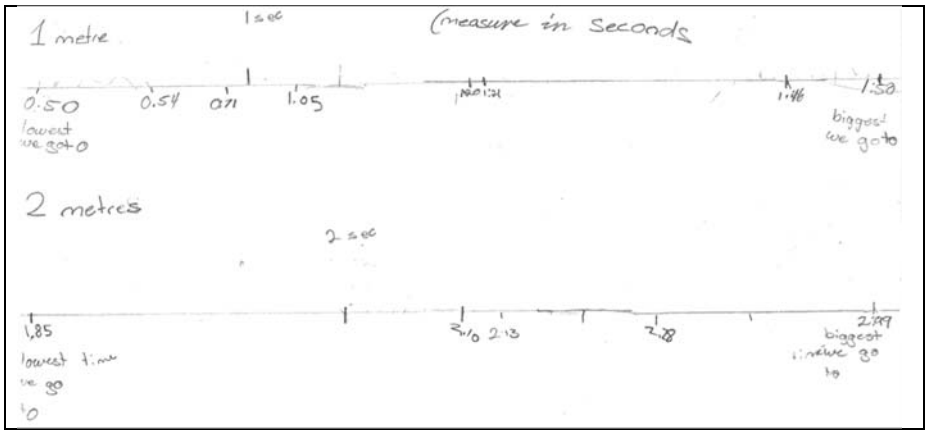


Figure 8. Examples of dot plots that marked flight times in rank order

In summary, although the majority of the groups created pairs of dot plots that could be used as evidence to argue whether or not flight time was related to drop height, some of the representations made the comparison easier than others (e.g., those shown in Figure 5). Some groups continued to use a single number line for their dot plot, labeling the dots to convey additional information about drop height. This variation in dot plot representations and the idea that some representations facilitated comparing groups better than others set the stage for the next part of the MXA.

3.4. MILESTONE 4: IDENTIFYING USEFUL CHARACTERISTICS OF REPRESENTATIONS FOR COMPARING GROUPS

The representations generated by the students for showing evidence that "drop height matters" led to a continuation of the MXA, focusing on the useful characteristics of the various representations created by the students. The next question that was posed by the teacher was "which ones of these [representations] is the easiest to be able to tell, and most convincing, that [drop] height matters?" This question aligns with Hestenes' (2010) three elements of models (simplicity, generalizability and applicability) and focused students' attention on interpreting their graphs and using them to draw "convincing" conclusions. This question also aligns with the principles put forward by Lesh et al. (2003) for engaging students in self-evaluation of their models and model generalizability. The teacher also asked "what are some of the properties of some of these [representations] that make that [height matters] easy to see?" Here, the teacher focused the students' attention on the salient characteristics of the graphs (such as using the same scale, and alignment of the scales; see Figure 5) that make a representation easier to interpret. In other words, a dual focus was made possible by variation in the representations that the students generated: (1) How do you interpret these representations to answer the question about whether drop height matters? (This drew the students' attention to making sense of what they personally did and how they chose to represent their own data); (2) How can we evaluate the usefulness of the representations that were created by the class as a whole? By engaging students in this evaluative phase, the teacher provided the broad guidance necessary for students to explain what characteristics of the representation (e.g., same scale for both drop heights, labeling the drop height for each graph, consideration of the number of data values shown on the graph) are useful in making conclusions based on a comparison of two different drop heights.

After all of the representations were discussed, the teacher focused on the representations of the two groups that each had two vertically aligned number lines on the same scale and with equal intervals (shown in Figure 5). At this point, the teacher asked:

> So, if we were going to sit back and look at all of these, and see which ones of these is the easiest to be able to tell, and most convincing that height matters, what are some of the properties of some of these that make that easy to see?

In reference to the group's representations shown in Figure 5b, the teacher asked questions such as, "Can you see anything that makes that one easier to interpret? What is it that that group has done that makes it clear to you?" One student noted that the clusters of points between the two drop heights were not overlapping. When the teacher asked what the group did that allowed the class to see the distributions were not overlapping, another student stated that the same scale was used for both dot plots. The teacher then drew students' attention to other groups' representations where the scales were not the same for the two drop heights (e.g., see Figures 6, 7, and 8) and asked the students to contrast the ease of judging the difference. For example, in discussing the dot plot in Figure 6:

Teacher:	And if you're standing in the back of the room, is that something you can see right away from these two? That these ones [pointing to 2-metre dot plot] took longer than these ones [pointing two 1-metre dot plot]? What makes it hard, Martin? You're shaking your head.
Martin:	Um, well, maybe it's, um, we did the first one from 0 to 2 and then we did the second one from 1 to 2, so they're both kind of in the middle.
Teacher:	OK.
Martin:	Except the second one actually got more time [speaking too soft to transcribe]
Teacher:	So, it is a little bit hard when you have different scales on your number lines to see that immediately. You have to really pay attention to the numbers. Is there another observation that you have about this data, by the way, besides just the middle? Anything else you notice about it? Simon, you're sort of mouthing to yourself.
Simon:	It's kind of all spread out.
Teacher:	This [2-meter dot plot] is kind of more spread out than that one. Do you see that? The scale makes it a little difficult to see, because if you had this [2-meter dot plot] go from 0 to 2 it would have been a bit more squished up [squeezing gesture with index finger and thumb]. But what does it tell you when it's more spread out? What does it tell you about those two flights? What does spread out mean [shrugging shoulders]? Ethan.
Ethan:	They're far apart, they're not close.

By the end of the model exploration portion of the model development sequence, the students' models (or systems) for comparing groups now included the use of paired dot plot representations. They had identified useful characteristics of representations for comparing groups such as using equal scales on both dot plots and the vertical alignment of the scales.

3.5. MILESTONE 5: APPLYING THE STUDENTS' MODEL FOR COMPARING GROUPS

The last portion of the model development sequence was an MAA where the students had to apply the model for comparing groups that was elicited and explore a new inquiry question: How big of a difference is there between the flight duration of short-wing and long-wing paper helicopters? Short-wing versions of the helicopters were made by trimming the wings at the vertical line marked on the wings in Figure 2. The 24 students who attended this lesson worked primarily in pairs, with one group of three students and

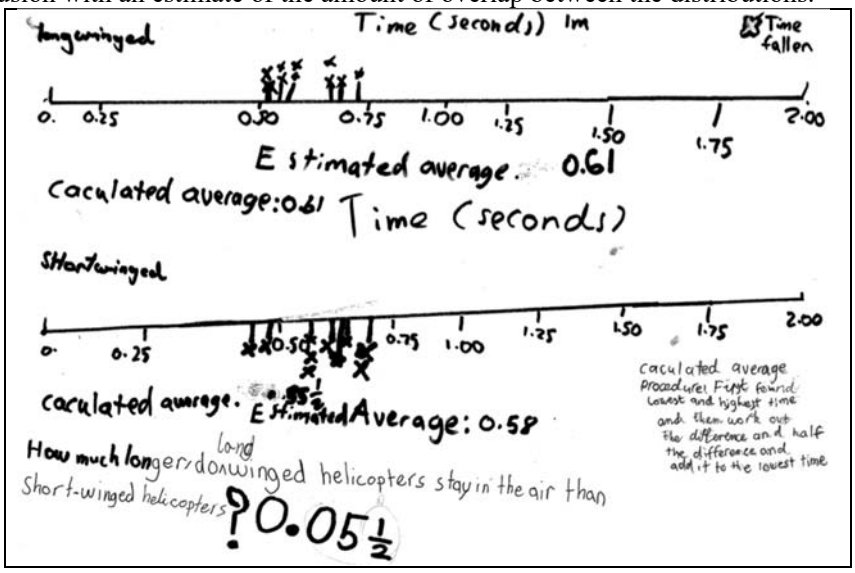
one student who worked alone. All of the groups attended to drop height by selecting a fixed drop height (most commonly of one meter) for dropping their long-wing and short-wing helicopters. All of the groups created graphical representations that consisted of two horizontal number lines, one for each type of helicopter, that were of the same physical width, used the same scale, and were aligned vertically one above the other.

These representations furthered the development of students' models for the comparison of groups as they now had to account for the overlap in the distributions and draw conclusions based on the extent of the overlap. The distributions of flight times for the short-wing and long-wing helicopters overlapped almost completely for six of the groups (see Figure 9), with only a small portion of the two distributions overlapping for another three groups (Figure 10), and complete separation of the distributions for the remaining three groups (Figure 11).

Only two of the groups that had almost complete overlap of the two distributions acknowledged the overlap in their conclusion statement ("Some of the results were even the same"; "a lot of other short winged ones are bigger than the smallest long winged helicopter"). Both of these groups stated there was no difference between the flight times of the two helicopter types ("no real big difference"; "there similar").

Of the four remaining groups with almost complete overlap in the distributions, one group (Figure 9a) concluded "that long-winged helicopters stay in the air roughly close to the short-winged helicopters by an estimate of 0.055." Another group (Figure 9b) estimated the average difference to be "around 0.10 second", but also stated "which isn't long." Of the two remaining groups, one simply stated a very small estimated average difference (2/100 of a second) but without commenting if that represented a meaningful difference. The last group (Figure 9c), contrary to all other groups, stated, "Our evidence shows us that the short-winged helicopter stays in the air the longest. The highest in the short winged is 1.95 seconds," even though two of the long-wing helicopter flight times were greater than 1.95 seconds. Overall, the five of the six groups with almost completely overlapping distributions of flight times concluded their evidence did not indicate a meaningful difference in flight times between the two types of helicopters, with two groups supporting this conclusion with an estimate of the amount of overlap between the distributions.

a.



b.

c.

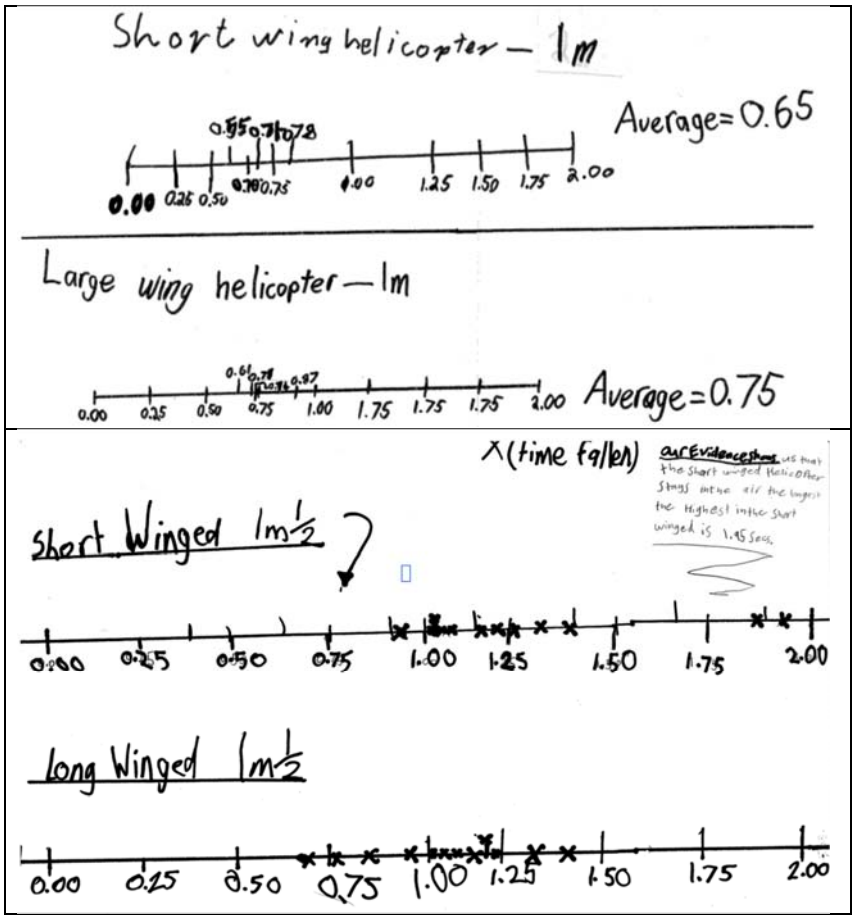


Figure 9. Examples of dot plots from the MAA with almost complete overlap in the distribution of short-wing and long-wing helicopters

As stated earlier, only a small portion of the distributions of flight times for the two helicopter types overlapped for three groups, and there was complete separation of the distributions for the remaining three groups. Five of these groups stated that their evidence "shows" or "proves" that the long-wing helicopters stayed in the air longer than the short-wing helicopters.

a.

b.

c.

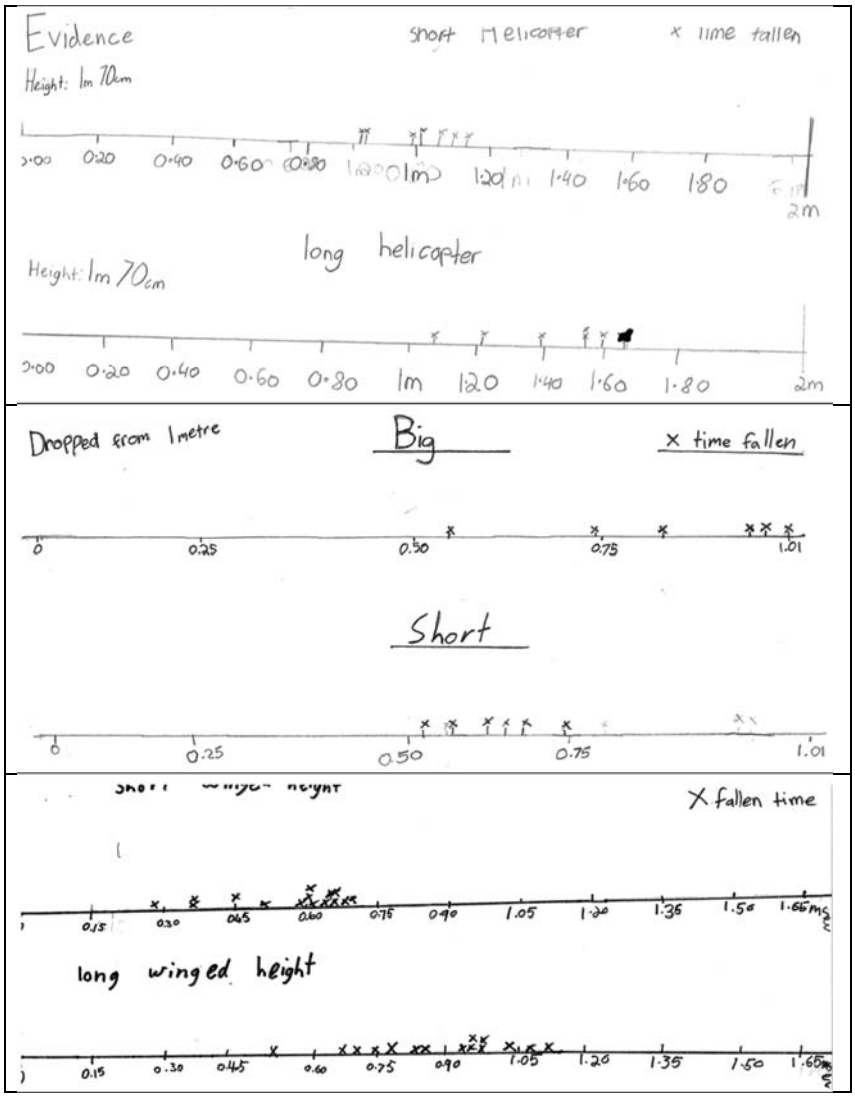


Figure 10. Examples of dot plots from the MAA with slight overlap in the distribution of short-wing and long-wing helicopters

a.

b.

c.

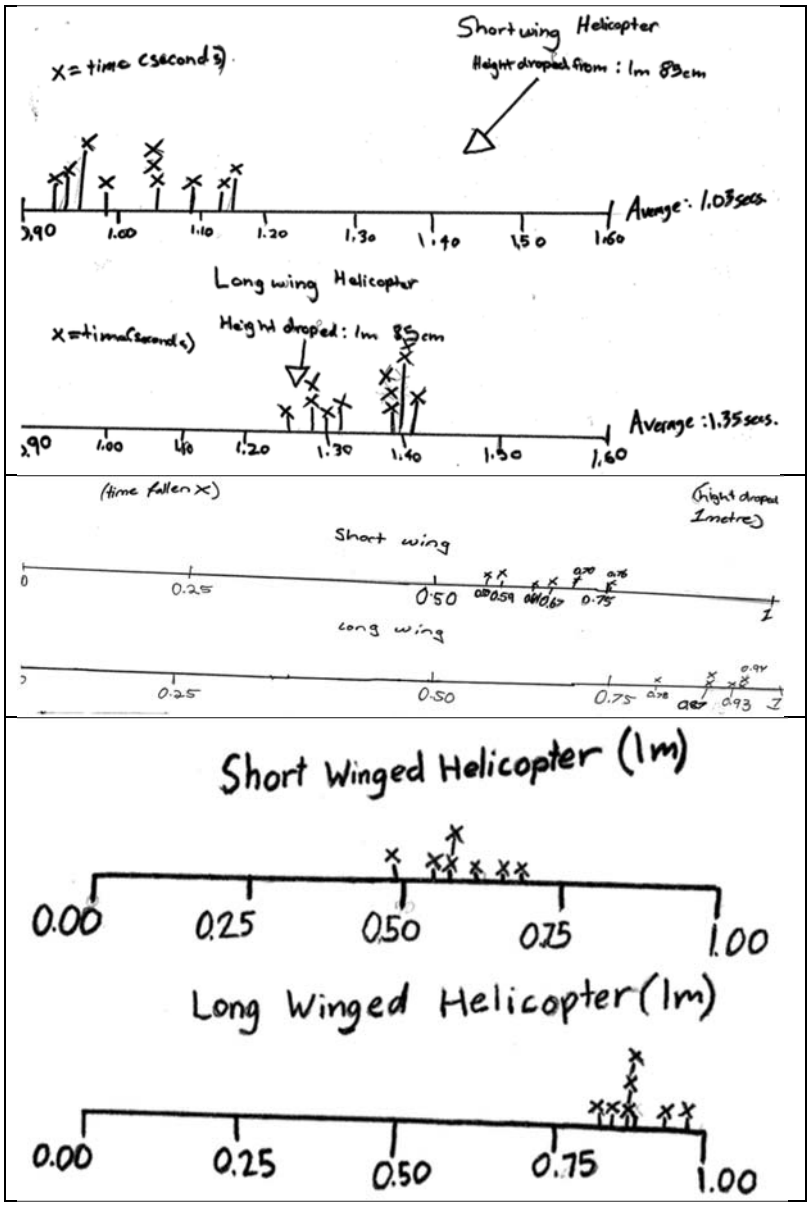


Figure 11. Examples of dot plots from the MAA with no overlap in the distribution of short-wing and long-wing helicopters

In addition to their graphical representations, most (9, or 75%) of the groups also reported an estimated difference between the flight time of the short-wing and long-wing helicopters. However, they had various ways of estimating and reporting the difference. Five of these groups reported an "average" value for each helicopter type and the reported difference equaled the difference between the two averages (note that students had not yet been taught how to calculate the mean). Three groups described how they estimated the averages: two groups calculated the mid-range value (i.e., the midpoint between the

minimum and maximum values) for each type of helicopter, whereas the third group "visualised the averages, from the dot plot." For the four groups that provided an estimated difference but did not report an average for each type of helicopter, one group reported the difference between the maximum flight times for each type of helicopter. For the other three groups, the estimated difference was close to the difference between the two mid-range values. We note that for some groups of students, when doing calculations, they tended to return to single-value estimates of the difference in flight times for short-wing and long-wing helicopters.

4. DISCUSSION

In this article, we have addressed the following research question: How did the model development sequence support students in developing a generalizable model for comparing two groups of observed data? We respond to this question by focusing on three aspects of the question: (1) the role that the model development sequence had in supporting students; (2) the development of students' models over the three lessons; and (3) the generalizability of the students' models for comparing two groups.

4.1. THE ROLE OF THE MODEL DEVELOPMENT SEQUENCE

The lessons were designed around Lesh and colleagues' (Lesh et al., 2003) model development sequence of an MEA, an MXA, and an MAA. This model development sequence provided an instructional design framework that supported students through a process of needs-based revision and refinement of their models. Our application of the model development sequence built on the modeling perspectives of Hestenes (2010), Lehrer and Schauble (2010), and Barbosa (2006) to promote development of a model that could productively be applied and re-used across a range of contexts. The model development sequence guided the research team to design a meaningful task that elicited a need for a process or system for comparing two groups (MEA), then explored and improved on representations that compared two groups (MXA). Finally, the application of students' models to a new situation with a similar structure (MAA) further strengthened the generalizability and abstraction of the model.

The model development sequence provided a focus and goal for each lesson. In each stage of the sequence, the lessons were revised to adapt to outcomes of the prior lesson and capitalize on opportunities that emerged. Collectively, the model development sequence provided us with an instructional design process to elicit, explore, and apply a sequence of improvable models that were authentically integrated with students' experiences. The end goal of the sequence was to support students to create shareable and generalizable models for comparing two groups. The sections below further elaborate on this goal.

4.2. THE DEVELOPMENT OF STUDENTS' MODELS

In Section 3, we described the details of how the students' models developed over the three lessons. Students' models began from their initial ideas about the context they were investigating: How long does a paper helicopter stay in the air? To address this question, students grappled with a number of contextual factors that would challenge a direct response. Wrestling with these factors fueled discussion and allowed students multiple opportunities to scrutinize their ideas and improve their models. In this way, the development of students' models went through a number of milestones in which new insights were gained that provoked a need for model improvements.

The students' initial work with flying the helicopters in the MEA appeared as unstructured and unsystematic play, with a large variety of techniques used to launch the paper helicopters. This "play" revealed a number of important insights for students that assisted them to select aspects of the context to attend to for the initial model that mattered to them (or to ignore, if it did not) (Nemirovsky & Tierney, 2001), such as the need to measure and record drop times, the need to consider variability and its possible sources, and noticing (conjecturing) that drop height affected flight time.

Initially to answer the problem they were addressing, the teacher asked students to record their group's data on a single number line for the whole class. This initial representation of data (Figure 4) provided a first common artifact for the class to interpret and critique (Moschkovich, 2008). The students' play in the MEA gave them experiences to connect the context of their helicopter flights to the data on the number line. The experiences, connections, and insights that emerged from students' play and classroom discussion promoted a need for further development of their models. Throughout this and later milestones, students had to identify and articulate their thinking, make meaning of the specifics and abstractions of issues that arose (e.g., think about an appropriate sample size), test ideas, contend with conflicting information, attend to evidence, and make decisions based on their model. In the MXA, students used what they learned from the MEA to develop improved models that included useful representational structures for sharing their models of the helicopter phenomena (Figures 5-8). These representations acted as common artifacts to promote fruitful discussion and negotiation around theoretically useful and personally meaningful aspects of each model (Barbosa, 2006; Lehrer & Schauble, 2010; Moschkovich, 2008). In this way, the models students produced in the MAA in the third lesson were more likely to adapt aspects of previous models that were seen as productive for responding to a new problem situation (Doerr & English, 2003; Lehrer & Schauble, 2010). Therefore, the children's models were situated in and shaped by the context of their activities (Font et al., 2007; Roth & McGinn, 1998).

The teacher played a key role in each lesson activity by encouraging students to play and make sense of the flight of paper helicopters, introducing representations (a common number line for class data), asking students to interpret and critique representations (e.g., "Drop height seems to matter. How can we tell this from the class data?"), selecting representations to compare, asking questions about characteristics of useful representations (e.g., "what are some of the properties of some of these [representations] that make that [height matters] easy to see?"), and privileging aspects of each discussion that promoted the overall goals of the sequence of activities (e.g., "Can you see anything that makes that one easier to interpret? What is it that that group has done that makes it clear to you?"), thus structuring and guiding the overall development of students' models. These aspects of the teacher's role highlight two important instructional implications for using a modeling approach. First, as seen in the unstructured play during the MEA and in the openness of the initial question for investigation, the students were free to approach the problem situation in ways that made sense for them. This generated a multiplicity of approaches as students engaged in deciding how and what data to collect, in interpreting those data, and in critiquing the usefulness of various representations of the data. Second, the multiplicity of student approaches provided the teacher and, more importantly, the students, with opportunities to compare representations and their usefulness in conveying information to others (how is "that one easier to interpret?") and in providing evidence for conclusions ("how can we tell?"). Thus, the modeling approach engaged students in the critical activity of the evaluation of the goodness and usefulness of their models.

4.3. STUDENTS' GENERALIZABLE MODELS FOR COMPARING GROUPS

Our research question included a third aspect—a generalizable model—that ensured an outcome for students that would deepen their theoretical knowledge of statistics and facility with data-based models (appropriate to their age). Generalizable models represent key structures of a discipline that appear across multiple contexts. In this study, the generalizable model was that of comparing two groups. Statistical situations that can be modeled by comparing two groups are common. The students' journey through the model development sequence provided them with a useful model for comparing two groups. The sequence ended with the application of their specific and situated model being adapted and re-used for a different type of group comparison in the same context. With additional experiences and instructional support for making group comparisons in new contexts, it would be anticipated that students would be more likely to recognize a situation that called for group comparisons and have a general model at hand that they could modify and apply to the situation. Transfer of problem solving knowledge and skills to novel contexts, however, is known to be difficult to achieve and dependent on many factors, such as the development of rich schema, intentional application of learned methods in new contexts, and large amounts of experience (e.g., Bransford, Brown, & Cocking, 2000; Lovett & Greenhouse, 2000; Singley & Anderson, 1989). Research is needed on conditions that promote effective transfer of models developed through model development sequences.

Having a target of a generalizable model—an outcome of the model development sequence—provides a systematic approach for statistics educators to design learning experiences that address the key structures of the discipline. Rather than isolate concepts and skills such as graph construction and interpretation, calculation of center and spread, and managing samples and sampling, a models and modeling approach that includes generalizable (and *useable*) models places an emphasis on statistics as both powerful and relevant. The journey towards a generalizable model included multiple discussions about the useful characteristics of representations that were authentic to the students for addressing the problem under investigation. These explicit conversations were critical in facilitating the students' journey and thus point to an important implication for instruction based on model development sequences. The instructional goal of such sequences is not the particular solution to a particular problem situation. Rather the goal of instruction is for the development of a model that can be used and re-used in a range of structurally similar problem situations. Hence, the design of activities and the concomitant role of the teacher is to make visible the use and re-use of the model through its initial creation (MEA), representation (MXA), and application (MAA). In other words, a modeling approach to instruction has at its core the design of tasks and classroom activity that is focused on generalizability and applicability as instructional goals.

4.4. ADDRESSING KEY CHALLENGES IN STATISTICS EDUCATION

We end our discussion by returning to the issues raised in the introduction, that of addressing Bakker and Derry's (2011) three key challenges in statistics education: avoiding inert knowledge, avoiding atomistic approaches and fostering coherence in students' learning experiences, and sequencing topics to improve this coherence for students. We argued that research in statistics education around informal inferential reasoning has moved the field forward on the second challenge. That is, research over the past decade has provided numerous examples of how informal inferential reasoning provides a more holistic approach to curriculum, thus fostering coherence. Therefore, informal inferential reasoning holds tremendous potential for improving the teaching and learning of statistics. On its own, however, informal inferential reasoning does not provide sufficient guidance

to instructors on how to support students in applying their statistical knowledge to novel problems nor how to sequence topics to improve coherence.

By building on research about informal inferential reasoning with a perspective of model-based reasoning (Lehrer & Schauble, 2010), we see opportunities for addressing Bakker and Derry's two remaining challenges. The first challenge—avoiding inert knowledge to allow students to apply what they know to novel problems—is addressed by modeling because the focus of modeling is meaningful application that captures generalizable structure of a phenomenon (Hestenes, 2010). The third challenge put forth by Bakker and Derry—sequencing topics for improved coherence for students—is addressed by using model development sequences (Lesh et al., 2003) to organize instruction around eliciting students' initial conceptions about their experiences (an MEA), exploring useful and productive representations of their concepts (an MXA), and applying and modifying their models in new situations (an MAA). Working with a generalizable structure of a phenomenon—as the children in our study did in comparing two groups—supports them to recognize isomorphic elements within novel situations having the same (or similar) structure. The final problem in our study—seeking comparisons between the flights of short-wing and long-wing helicopters—was similar to the problem of comparing helicopter flights dropped at different heights. The common structure—comparing two groups of observed data—enabled the students to work with the isomorphic elements from the problem of comparing times from helicopters with different wingspans—stacked distributions, common scale, collecting enough data, measuring systematically—that had served them in the problem of comparing times from different drop heights.

The models and modeling perspective used in this study integrated informal inferential reasoning as a natural aspect of modeling. If informal statistical inference is seen as a holistic approach for integrating topics in statistics education (Bakker & Derry, 2011), then modeling shifts the notion of sequencing *topics* to sequencing students' modeling *experiences*. Researchers have illustrated how students' experiences with informal statistical inference and modeling, even at a young age (Fielding-Wells & Makar, 2015; Makar, 2016; Paparistodemou & Meletiou-Mavrotheris, 2008), overlap and build key topics over time rather than sequence them. The model development sequence—MEA, MXA, MAA—developed by Lesh, Doerr, and their colleagues (Doerr & English, 2003; Lesh et al., 2003) that was used in this study provided a framework for meaningfully sequencing students' instructional experiences.

ACKNOWLEDGEMENTS

The data from this paper come from a grant funded by the Australian Research Council (DP120100690). The first and second authors' airfares to Australia were paid by Travel Awards for International Collaboration Research (Category 1) from The University of Queensland.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum.

Ainley, J., Pratt, D., & Nardi, E. (2001). Normalising: Children's activity to construct meanings for trend. *Educational Studies in Mathematics, 45*(1-3), 131-146. doi:10.1023/A:1013822512833

Annis, D. H. (2005). Rethinking the paper helicopter. *The American Statistician, 59*(4), 320-326. Retrieved from http://dx.doi.org.ezp3.lib.umn.edu/10.1198/000313005X70777

Ärlebäck, J. B., Doerr, H. M., & O'Neil, A. H. (2013). A modeling perspective on interpreting rates of change in context. *Mathematical Thinking and Learning, 15*(4), 314-336. Retrieved from http://dx.doi.org.ezp3.lib.umn.edu/10.1080/10986065.2013.834405

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning, 13*(1-2), 5-26. http://dx.doi.org.ezp3.lib.umn.edu/10.1080/10986065.2011.538293

Barbosa, J. C. (2006). Mathematical modelling in classroom: A socio-critical and discursive perspective. *ZDM, 38*(3), 293-301. doi:10.1007/BF02652812

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education* (Proceedings of the 7th International Conference on the Teaching of Statistics, Salvador, Bahai, Brazil, July 2-7). Voorburg, The Netherlands: International Association for Statistical Education and the International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots7/2D1_BENZ.pdf

Ben-Zvi, D., & Garfield, J. (2004). *The challenge of developing statistical literacy, reasoning, and thinking*. Dordrecht, The Netherlands: Kluwer.

Box, G. (1992). Teaching engineers experimental design with a paper helicopter. *Quality Engineering, 4*(3), 453-459.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: National Academies.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13. Retrieved from http://www.jstor.org.ezp3.lib.umn.edu/stable/3699928

delMas, R., Garfield, J., & Zieffler, A. (2014). Using TinkerPlots™ to develop tertiary students' statistical thinking in a modeling-based introductory statistics class. In T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit Werkzeugen, Mathematik und Stochastik lernen - Using tools for learning mathematics and statistics* (pp. 405-420). Wiesbaden, Germany: Springer Spektrum.

Doerr, H. M., & English, L. D. (2003). A modeling perspective on students' mathematical reasoning about data. *Journal for Research in Mathematics Education*, *34*(2), 110-136. doi: 10.2307/30034902

English, L. D. (2012). Data modeling with first-grade students. *Educational Studies in Mathematics, 81*(1), 15-30. doi:10.1007/s10649-011-9377-3

Erhardt, B. E. (2007). Designing a better paper helicopter using response surface methodology. *STATS: The Magazine for Students of Statistics, 48*, 14-21.

Fielding-Wells, J., & Makar, K. (2015). Inferring to a model: Using inquiry-based argumentation to challenge young children's expectations of equally likely outcomes. In A. Zieffler & E. Fry (Eds.), *Reasoning about uncertainty: Learning and teaching informal inferential reasoning* (pp. 1-27). Minneapolis, MN: Catalyst Press.

Font, V., Godino, J. D., & D'Amore, B. (2007). An onto-semiotic approach to representations in mathematics education. *For the Learning of Mathematics*, *27*(2), 2-7, 14. Retrieved from http://www.jstor.org.ezp1.lib.umn.edu/stable/40248564

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education, 10*(3). Retrieved from http://www.amstat.org/publications/jse/v10n3/garfield.html

Garfield, J., Le, L., Zieffler, A., & Ben-Zvi, D. (2015). Developing students' reasoning about samples and sampling variability as a path to expert statistical thinking. *Educational Studies in Mathematics, 88*(3), 327-342. doi:10.1007/s10649-014-9541-7

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Gorard, S., Roberts, K., & Taylor, C. (2004). What kind of creature is a design experiment? *British Educational Research Journal, 30*(4), 577-590. doi:10.1080/0141192042000237248

Hestenes, D. (2010). Modeling theory for math and science education. In R. Lesh, P. L. Galbraith, C. R. Haines, & A. Hurford (Eds.), *Modeling students' mathematical modeling competencies: ICTMA 13* (pp. 13-41). New York: Springer.

Hovan, D., & Palalas, A. (2011). (Re)Conceptualizing design approaches for mobile language learning. *Computer-Assisted Language Instruction Consortium (CALICO) Journal, 28*(3), 699-720.

Kelly, A. E., Baek, J. Y., Lesh, R. A., & Bannan-Ritland, B. (2008). Enabling innovations in education and systematizing their impact. In A. E. Kelly, R. A. Lesh, & J. Y. Baek (Eds.), *Handbook of design research methods in education: Innovations in science, technology, engineering and mathematics learning and teaching* (pp. 3-18). New York: Routledge.

Konold, C., & Harradine, A. (2014). Contexts for highlighting signal and noise. In T. Wassong, D. Frischemeier, P. R. Fischer, R. Hochmuth, & P. Bender (Eds.), *Mit Werkzeugen, Mathematik und Stochastik lernen - Using tools for learning mathematics* (pp. 237-250). Wiesbaden, Germany: Springer Spektrum.

Lehrer, R., Jones, R. S., & Kim, M. J. (2014). Model-based informal inference. In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education* (Proceedings of the 9th International Conference on the Teaching of Statistics, Flagstaff, Arizona, July 13-18). Voorburg, The Netherlands: International Statistical Institute. Retrieved from https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_6E1_LEHRER.pdf

Lehrer, R., Kim, M. J., & Jones, R. S. (2011). Developing conceptions of statistics by designing measures of distribution. *ZDM, 43*(5), 723-736. doi:10.1007/s11858-011-0347-0

Lehrer, R., & Schauble, L. (2010). What kind of explanation is a model? In M. K. Stein & L. Kucan (Eds.), *Instructional explanation in the disciplines* (pp. 9-22). New York: Springer.

Lesh, R., Hoover, M., Hole, B., Kelly, A., & Post, T. (2000). Principles for developing thought revealing activities for students and teachers. In R. Lesh & A. Kelly (Eds.), *Handbook of research design in mathematics and science education* (pp. 591-645). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R. A., Cramer, K., Doerr, H. M., Post, T., & Zawojewski, J. S. (2003). Model development sequences. In R. A. Lesh & H. M. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching* (pp. 35-58). Mahwah, NJ: Lawrence Erlbaum Associates.

Lesh, R. A., & Doerr, H. M. (Eds.). (2003). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lovett, M. C., & Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *The American Statistician, 54*(3), 1-11. http://dx.doi.org/10.1080/00031305.2000.10474545

Makar, K. (2013). Predict! Teaching statistics using informal statistical inference. *Australian Mathematics Teacher, 69*(4), 34-40.

Makar, K. (2014). Young children's explorations of average through informal inferential reasoning. *Educational Studies in Mathematics, 86*(1), 61-78. doi:10.1007/s10649-013-9526-y

Makar, K. (2016). Developing young children's emergent inferential practices in statistics. *Mathematical Thinking and Learning, 18*(1), 1-24. Retrieved from http://dx.doi.org.ezp3.lib.umn.edu/10.1080/10986065.2016.1107820

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1-2), 152-173. Retrieved from http://dx.doi.org.ezp3.lib.umn.edu/10.1080/10986065.2011.538301

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

Moschkovich, J. N. (2008). "I went by twos, he went by one": Multiple interpretations of inscriptions as resources for mathematical discussions. *Journal of the Learning Sciences, 17*(4), 551-587. Retrieved from http://dx.doi.org.ezp2.lib.umn.edu/10.1080/10508400802395077

Nemirovsky, R., & Tierney, C. (2001). Children creating ways to represent changing situations: On the development of homogeneous spaces. *Educational Studies in Mathematics*, *45*(1-3), 67-102. doi:10.1023/A:1013806228763

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal, 7*(2), 83-106. Retrieved from http://iase-web.org/documents/SERJ/SERJ7(2)_Paparistodemou.pdf

Peters, S. (2011). Robust understanding of statistical variation. *Statistics Education Research Journal, 10*(1), 52-88. Retrieved from http://iase-web.org/documents/SERJ/SERJ10(1)_Peters.pdf

Powell, A. B., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *Journal of Mathematical Behavior, 22*, 405-435. Retrieved from http://dx.doi.org.ezp2.lib.umn.edu/10.1016/j.jmathb.2003.09.002

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

Roth, W.-M., & McGinn, M. K. (1998). Inscriptions: Toward a theory of representing as social practice. *Review of Educational Research, 68*(1), 35-59. Retrieved from http://www.jstor.org.ezp1.lib.umn.edu/stable/1170689

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Siorek, T. L., & Hafta, R. T. (1998). Paper helicopter: Experimental optimum engineering design classroom problem. *Proceedings of the 7TH AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis and optimization*. Retrieved from http://arc.aiaa.org/doi/pdf/10.2514/6.1998-4963

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. New York: Routledge.

Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40-58. Retrieved from http://iase-web.org/documents/SERJ/SERJ7(2)_Zieffler.pdf

HELEN M. DOERR
215 Carnegie Hall
Syracuse University
Syracuse, NY 13244-1150