

EXPLORING BEGINNING INFERENCE WITH NOVICE GRADE 7 STUDENTS

JANE M. WATSON
University of Tasmania
Jane.Watson@utas.edu.au

ABSTRACT

This study documented efforts to facilitate ideas of beginning inference in novice grade 7 students. A design experiment allowed modified teaching opportunities in light of observation of components of a framework adapted from that developed by Pfannkuch for teaching informal inference with box plots. Box plots were replaced by hat plots, a feature available with the software TinkerPlots™. Data in TinkerPlots files were analyzed on four occasions and observed responses to tasks were categorized using a hierarchical model. The observed outcomes provided evidence of change in students' appreciation of beginning inference over the four sessions. Suggestions for change are made for the use of the framework in association with the intervention and the software to enhance understanding of beginning inference.

Keywords: *Statistics education research; Hat plots; Informal inference; Middle school students; TinkerPlots*

1. LITERATURE REVIEW

The word “inference” can convey many shades of meaning, depending on the context in which it is used and the adjectives that may be placed in front of it. A dictionary such as *Chambers* (Kirkpatrick, 1983) suggests an inference is “that which is inferred or deduced: the act of drawing a conclusion from premises: consequence: conclusion” (p. 644). For a statistician it is the premises that are important. For David Moore (1991) statistical inference is the “formal methods for drawing conclusions from data taking into account the effects of randomization and other chance variation” (p. 330). The parts of Moore’s definition that preclude it from most of the school curriculum are the “formal” methods, the “randomization,” and the “chance” variation to the extent that it relies on formal probability. Within the statistics education community, the phrase “formal inference” is usually used synonymously with “statistical inference” in Moore’s sense. This usage leads to attempts to define “informal inference.” Certainly at the middle school level replacing “formal methods” with “informal methods” is appropriate as the mathematics required for formal methods is not available to students. The question of what constitute appropriate informal methods is then a matter for debate among statistics educators. Although randomization may be introduced to middle school students through chance devices such as coins and dice, the link to the need for random methods generally may be difficult for students to grasp and even more difficult to implement during data collection activities. Chance variation is an idea that is accessible in an intuitive fashion but nearly impossible to connect to numerical values that would express degrees of uncertainty.

This raises the question, *What aspects of the process of drawing conclusions from data can be kept and what left out, and still use the phrase “informal inference”?* Rubin, Hammerman, and Konold (2006) give their summary of important ingredients of informal inference as including properties of aggregates related to centers and variation, sample size, controlling for bias through sampling method, and tendency (acknowledging uncertainty). Watson and Moritz (1999) use the phrase “beginning inference” to describe the comparison of two finite data sets because the activity is a precursor to the use of t tests by statisticians. The context is similar but the premises differ. Although the techniques available fall under the heading of exploratory data analysis, the requirement to make a decision about the difference between the two groups takes the activity into the realm of beginning or informal inference. The admission of uncertainty about the decision reached, although based on intuition and maybe graphs rather than numbers, mirrors the process involved with formal inference.

The National Council of Teachers of Mathematics’ (NCTM) *Principles and Standards for School Mathematics* (2000) use the words “develop and evaluate inferences and predictions” in the Data Analysis and Probability Standard throughout the document for all ages, with increasingly sophisticated and rigorous criteria used for the techniques that “are based on data” (pp. 48-51). Hence, the use of the word “inference” on its own does not necessarily imply “formal inference.” Although they do not use the phrase, the descriptions of the NCTM would imply agreement with the definition of “informal inferential reasoning” by Ben-Zvi, Gil, and Apel (2007):

Informal Inferential Reasoning (IIR) refers to the cognitive activities involved in informally drawing conclusions or making predictions about “some wider universe” from patterns, representations, statistical measures and statistical models of random samples, while attending to the strength and limitations of the sampling and the drawn inferences. (p. 2)

There are many components in this definition and it is questionable whether a unit on beginning inference in most middle school programs would be able to cover them with a high level of sophistication.

The context of the study reported here included the provision and introduction of the software TinkerPlots™ (Konold & Miller, 2005). This allowed for the consideration of the use of the software in relation to the goals of informal inference. TinkerPlots was designed using a bottom-up constructivist approach for the development of students’ data handling and analysis skills (Konold, 2007). Students build representations using tools and the drag-and-drop facility, rather than choosing a “type” of plot that is ready-made. One particular tool of relevance to this research is the hat plot. It is a simpler version of the box plot. In its basic default form the hat consists of a crown that covers the middle 50% of the data with brims on either side covering the 25% of data at each extreme. A hat plot for a small data set is shown in Figure 1. The median is not represented in the hat plot although it can be marked on the horizontal axis where the data are stacked. A hat plot initially appears in conjunction with the data set it represents, providing visual reinforcement.

The similarity of a hat plot to a box plot suggests that the research of Pfannkuch (2005, 2006a, 2006b) on informal inference with box plots in the final years of school may be instructive for analyzing the data in this study. Pfannkuch devised a model for the development of inferential reasoning based on box plots. The model evolved from Pfannkuch’s intensive planning with a classroom teacher before teaching informal inference based on box plot representations, observations in the classroom during teaching, analysis of the data, and various discussions with statisticians and statistics educators. The eight elements of reasoning in the model are (i) *hypothesis generation*,

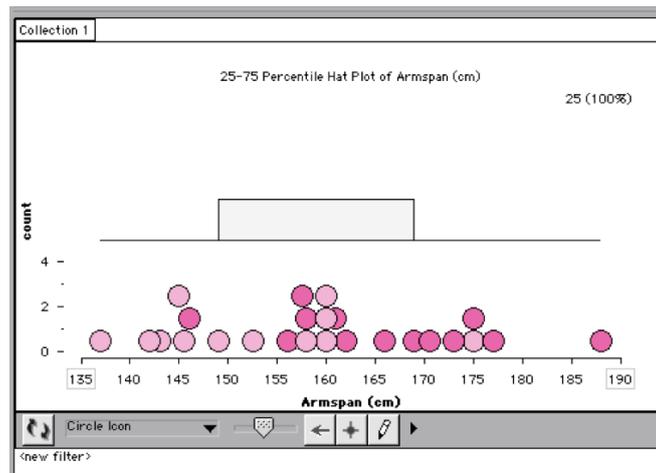


Figure 1. Hat Plot for a data set

(ii) *summary* of the plot, (iii) considering *shift* of box plots, (iv) comparing the *signal* in the central 50% of data, (v) considering *spread*, (vi) discussing aspects of *sampling*, (vii) considering *explanatory* aspects of the *context*, and (viii) considering *individual cases* (Pfannkuch, 2006b). Two other moderating elements were also described relating to weighing up *evidence* for evaluation and to using concept *referents* that lay behind the box plots, which were the objects of the lessons. As these were considered by Pfannkuch to be included in the main eight elements (2006b, p. 33), they are not included in this study. Evaluation and adaptation of the model was based on observations of teachers and their students in classrooms, assessment tasks, and associated student outcomes. One of the difficulties observed in the classroom was the abstract nature of the box plot, which was presented without the display of data values, putting extra pressure on students to recall the proportional reasoning associated with interpreting the plot. Pfannkuch's conclusions recommended that the data and box plot should appear together.

In relation to the use of box plots, Bakker's (2004) research on using computer tools with middle school students led to the questioning of the introduction of box plots to middle school students (Bakker, Biehler, & Konold, 2005). The features of box plots that cause difficulties for students include (i) the lack of ability to examine individual values, (ii) the different form of presentation from other graphical representations, (iii) the non-intuitive nature of the median, and (iv) the visual presentation of density rather than frequency. Based on the five-number summary (minimum, maximum, median, and first and third quartiles) the box plot is an intermediate summary of data between a graph and a single summary statistic such as the arithmetic mean. As such it may hide some interesting features of a data set. Bakker et al. recommended that the introduction of box plots be delayed until later in the middle school years and that more time be spent relating box plots to the data sets they represent. These recommendations support the possibility that hat plots as available in TinkerPlots may provide a viable alternative to box plots for middle school students. Other recent research using TinkerPlots with both students (Friel, O'Connor, & Mamer, 2006) and teachers (Rubin & Hammerman, 2006) illustrates aspects of Pfannkuch's model adapted for use with the different visual presentations provided by TinkerPlots. Friel et al. note the use of dividers, box plots, and reference lines to assist students in looking at data sets as aggregates in much the same way that hat plots are used in this study. Rubin and Hammerman use bins and dividers to consider slices of data sets for similar purposes. The significant aspect of visualization in both

studies, enforcing attention to distribution, highlights the four elements of Pfannkuch's model related to Spread, Shift, Signal, and Summary. Using a framework for assessing enabling aspects of software applied to statistical environments, Fitzallen (2007) found that TinkerPlots (i) was accessible and easy to use, (ii) assisted recall of knowledge and allowed data to be represented in multiple forms, (iii) facilitated transfer between mathematical expression and natural language, (iv) provided extended memory when organizing or reorganising data, (v) provided multiple entry points for abstraction of concepts, and (vi) provided visual representations for both interpretation and expression (p. 24).

In relation to the model for observing the development of inferential reasoning with box plots by Pfannkuch (2006b), Konold et al. (2002) reported on observations of student descriptions of stacked dot plots that focused on "modal clumps." These clumps of data values occupied central locations in data sets and usually divided the distribution of values into three parts. The percentage of values in each part depended on the overall shape of the distribution. It appeared natural for the students to "see" the distributions in three parts: low, middle, and high. The modal clumps were often described in conjunction with a generalized average that was wider than a specific value. These observations also link to the display presented by a hat plot, with the natural splitting of the data set into three parts. Whether these possibilities for clustering assist in making decisions about differences in two data sets has received little research attention. Cobb (1999), however, observed students in a classroom study begin to consider the "hills" apparent in two data sets and how these might determine a difference in two groups. Watson and Moritz (1999) also found students using shape to distinguish between two classes' test results where the classes were of a different size.

2. CONTEXT AND RESEARCH QUESTIONS

The context within which the opportunity to conduct this study arose was a larger research project providing professional learning in mathematics for rural middle school teachers in an Australian state. The professional learning was aimed at providing teachers with the chance to plan and develop skills to assist their students with the mathematical foundations both for quantitative literacy as a life skill and for moving on to higher mathematics courses. The schools in the project were considered culturally isolated from the capital city and their students were performing on average below the national benchmarks for literacy and numeracy. TinkerPlots software was provided as part of the project to all schools and hence some professional learning sessions were based around it. Schools or teachers could elect to be involved with the researchers in case studies as part of the larger project. This provided the setting for addressing the following research questions:

- What are the observed learning outcomes for a class of grade 7 students, in relation to beginning inference in a learning environment supported by a learning and assessment framework, professional learning for the teacher, and a software package for data handling?
- How do the learning outcomes suggest changes to the framework, the implementation, or the interaction with software?

3. METHODOLOGY

The implementation of the study fits into a design experiment model of educational research as outlined for example by Cobb, Confrey, diSessa, Lehrer, and Schauble (2003)

and The Design-Based Research Collective (2003). Five criteria derived from these authors are satisfied by the study. First, the design intertwined the starting framework (or theory) for development of informal inference based on box plots with the classroom intervention employing TinkerPlots software that provided the hat plot as an interpretation tool. Second, the design was interactive in that four data collection sessions took place with adaptation to plans and analysis between sessions through ongoing discussion among the researcher, teacher, and a teacher-researcher (T-R). Third, the study was highly interventionist in the provision of a T-R to mould activity when the teacher appeared insecure and in the expectation of student participation in an environment that was unfamiliar to them (TinkerPlots). In terms of initial expectations there were both successes and failures that required adjustment. Fourth, a variety of data sources were connected to document the design and intervention, including student outcomes, with descriptive accounts and structural analysis of students' TinkerPlots output. Fifth, the modification to the teaching and classroom environment was related to the theoretical framework as a way to document student outcomes and make recommendations for future intervention.

To expand on these criteria in order to present the results in context, the following sub-sections outline the background in terms of the participants, including the preliminary professional learning of the classroom teacher, the basic context for each of the four teaching sessions, the nature of student output and its analysis in terms of the adapted Pfannkuch framework, and the other sources that contributed to the modifications made to theory and practice.

3.1. PARTICIPANTS' BACKGROUND

The study was based on a class of approximately 15 grade 7 students (aged 12-13 years) in a rural school (grades K to 10) in Australia. Not all of the students were present at all sessions. This was a convenience sample dictated by a larger research project. The class teacher, Jenny (pseudonym), was in charge of all sessions except the last. A research associate who was a trained teacher (the T-R), assisted the author in preliminary sessions with teachers and was present at all student classes, taking detailed notes, assisting with use of TinkerPlots, and occasionally teaching.

Before the four sessions where data were collected, Jenny attended a professional learning day with the author and T-R where TinkerPlots was introduced as a tool to assist in drawing informal inferences from data sets. The aims included

- (i) learning to enter data sets into TinkerPlots,
- (ii) using a student-created data set and TinkerPlots to discuss differences in data over time and how to determine if a "real" change has occurred,
- (iii) comparing data sets to decide if they are different,
- (iv) considering the issue of sample size when making a decision, and
- (v) becoming aware of the populations that samples represent.

Jenny supported other teachers experiencing similar activities and planned data collection and teaching sessions for her own students. Her aim was to address the five points.

3.2. LESSON CONTEXTS

Session 1 Prior to the session, pulse rates were collected from the middle-school students (grades 5 to 8) after a "Jump Rope for Heart" activity. Jenny's grade 7 students had explored TinkerPlots on their own earlier and the initial two-hour session with the heart-rate data was designed by Jenny and observed by the T-R. Jenny planned for the

students to enter their class data in the TinkerPlots data cards and then create plots to describe the variation in the observed pulse rates and what might be considered typical of the group. After a discussion of their observations and suggested reasons for differences (proposed inferences), the students were to measure their resting heart rates, record these for the class, and enter the data onto their existing data cards. The objective then was to create another plot in TinkerPlots and compare it with the first in order to speculate about differences in the two sets of data. The final aim was for the students to write a report in a text box in TinkerPlots about the inferences/conclusions drawn.

Between Sessions 1 and 2, the author, Jenny and the T-R discussed more specifically the aims of informal inference in relation to the elements of the adapted Pfannkuch model. Although the initial professional learning for Jenny and for the other teachers where Jenny assisted had covered this material generally (but not the eight specific points) it was felt necessary to be more explicit in discussing them. These are summarized in Table 1 as a Beginning Inference Framework. It was intended that Jenny be aware of them but she was not asked to “teach” them to her class explicitly as a list. She agreed that this was her understanding.

*Table 1. Eight elements of a Beginning Inference Framework
(adapted from Pfannkuch, 2006b)*

Element	Description
Hypothesis Generation	Reasons about trends (e.g., differences)
Summary	Summarizes the data using the graphs and averages produced in TinkerPlots
Shift	Compares one hat plot with the other/s referring to change (shift)
Signal	Refers to (and compares) information from the middle 50% of the data
Spread	Refers to (and compares) spread/densities locally and globally
Sampling	Considers sampling issues such as size and relation to population
Explanatory/Context	Shows understanding of context, whether findings make sense, and alternative explanations
Individual Case	Considers possible outliers and other interesting individual cases

Session 2 Some of the students assisted fellow middle-school students with a similar activity the day after Session 1, reinforcing their software skills. Jenny then devoted another day with her class to reviewing the heart-rate task and gaining an appreciation for the value of the TinkerPlots hat plot. This was to be a very structured session because of the students’ lack of background with hat plots and percentages generally. The aims reported by Jenny were to link these two ideas, as well as to focus on the ranges, means, medians and scales (horizontal axis) of the two heart-rate graphs. Again the plan was for students to record their observations about differences in the two sets of data using the new tools and reasons for these differences (informal inference).

Session 3 To reinforce the skills and understanding achieved, several weeks later Jenny collected data on arm-span length from all of the 58 middle-school students. This was organized by Jenny outside of the class time and she prepared a color-coded handout for students to use when entering their data. The plan for this session was jointly devised by Jenny and the T-R. The plan was to create a “class dictionary” of terms on the whiteboard to reinforce the previous sessions. The class would then decide on variables in the data set, enter the data into TinkerPlots, produce plots to explore the data, and make conjectures about any differences they observed. Again students were to write a

final report in a text box in TinkerPlots, explaining their evidence to support their conclusions (informal inferences) related to the arm-span data set.

Session 4 At the beginning of the following school year, approximately 3½ months after Session 3, the T-R returned to the school for a final session with the students, now in grade 8 with a different teacher. The T-R taught the entire session with another teacher-researcher observing and the classroom teacher assisting a student with special needs. Because of the time gap, the session was planned to start with a review of the arm-span measurement activity, including a review of the terms used such as *x*-axis, *y*-axis, bins, stacking, mean, median, hat plot, brim and crown, reference lines, dividers, percentages, range, spread, and clustering. The T-R created several TinkerPlots graphs from the data using a data show with students grouped around a table. Two new terms were to be introduced for specific discussion and definition: hypothesis and evidence.

Students were then each to be provided with the TinkerPlots file as shown in Figure 2 and asked to complete the questions presented in the text boxes. The data were from a data set provided with the TinkerPlots software (Child Development.tp) and the data set was reduced to case number, gender, height at age 2 years, height at age 9 years, and height at age 18 years. Three graphs were produced with hat plots and the questions were designed to encourage students to make and support informal inferences. After answering the questions presented in Figure 2, students were to be given the opportunity to explore the data set further by changing the graphs in any way they liked, especially if it would tell the story better.

3.3. OUTPUT AND ANALYSIS

The model suggested by Pfannkuch (2006b) to describe the aspects of informal inference associated with the introduction of box plots was adapted for both the teaching and the assessment of student work completed using TinkerPlots. The eight elements of the Beginning Inference Framework outlined in Table 1 cover well what would be expected in the context of a TinkerPlots exploration for middle-school students. Analysis was based on TinkerPlots files created by students during each session and saved at the ends of the sessions. The first step in the analysis was to determine which of the eight aspects of the Beginning Inference Framework were displayed in each response for each student. Grids were created with 8 cells for each of the 18 students who participated in at least one of the four sessions. A “tick” in a cell indicated that a graph was produced that appeared to address an aspect of Pfannkuch’s model but that no text was entered into a text box. Text in a cell provided a summary of text box comments made by the student. A clustering procedure (Miles & Huberman, 1994, p. 248) was used to group together comments related to the elements of informal inference displayed by each student at the end of each session.

At this point the clusters of elements of informal inference were assessed using the Structure of Observed Learning Outcomes (SOLO) model devised and employed by psychologists Biggs and Collis (1982) and mathematics educator Pegg (2002). In their developmental and assessment framework, it is the manner in which the elements are combined that shows increased sophistication and complexity in relation to achieving a desired outcome. In this study, the desired outcome was a description of an informal inference taking into account the components of the Beginning Inference Framework. As a simplified example, when first presented with a data set, a student may only be interested in observing her own place in the data set, or the largest and smallest values in the set. This would be considered a *unistructural*, single-element response within the

Cards
Table
Plot
Slider
Text

Name _____

Child Development

Attribute	Value	Unit
CaseNo	1	
Gender	F	
Height_2years	87.7	cm
Height_9years	133.4	cm
Height_18years	158.9	cm
<new attribute>		

Body measurements for 136 children at three different ages: 2 years, 9 years, and 18 years.

Attribute Description

CaseNo: Number assigned to child
Gender: Child's gender
Height_2: Age 2 height (cm)
Height_9: Age 9 height (cm)
Height_18: Age 18 height (cm)

Child Development

Consider the difference in heights between the males and females at age 2, at age 9 and at age 18. What changes do you see?

Child Development

How do the graphs help you decide if there is a difference between males and females at the three ages?

Child Development

What would you hypothesise about the differences?

How do the graphs support your hypothesis?

What questions would you ask about the data?

Figure 2. Tasks for Session 4

context of the overall expectation of a beginning informal inference task. Another student might look at a graph, plot the mean and report the average value for the data set. Again the response would be considered unistructural. A student who put these two elements together, or perhaps added another comment on spread, would likely be considered to produce a *multistructural* response, especially if the comments were made in a sequential fashion without being specifically related to each other. Taking a number of the elements and combining them to make a related coherent argument in terms of informal inference would be considered a *relational* response. For example, a relational response might

observe the shape of a data set, with clustering in one part, and spread in another, discuss an individual outlier, and compare these features with those of another data set hypothesizing a difference or sameness between them. No response or an idiosyncratic response is called *prestructural*.

3.4. OTHER SOURCES

The classroom interactions are described based upon the written notes of the T-R, long discussions of the researcher and T-R, and short meetings or telephone discussions with the teacher. One longer meeting of the teacher, T-R, and researcher was held after the first session. These sources are intertwined with evaluations of happenings in relation to the positive or negative impacts on the classroom atmosphere and student observed outcomes.

4. RESULTS

4.1. SESSION 1 – HEART-RATE DATA

Because students had explored TinkerPlots in a previous lesson, the expectation was that all students would open data cards to enter the data. After this there was, however, some confusion on the relative importance of name and heart rate and where to position them on the plots. Because of the way the data were entered on the TinkerPlots data cards, the names were paramount. Students fixed on these because they knew themselves and their classmates, and because “name” was the initial attribute on the data card. Almost all students created either a value plot or a plot with the same information on heart rates represented as circle icons. Some ordered these from lowest to highest heart rate and many noted the person with the highest rate and the person or persons with the lowest.

At this point the T-R intervened and asked all students to open a text box and write what their graphs told them about the heart rates. No one had yet stacked the dots on a dot plot. When asked about the “average” heart rate, most students looked at the data table and selected the most common value, 120 beats per minute. Some looked at highest and lowest values and commented on fitness. The resting heart rates were then measured. The T-R questioned the students about where the new data would go, what they were as attributes, and why the same data cards were used to record resting values. Then students had to make decisions about how to display the new data. When required, the T-R encouraged the students to create two graphs, one for each of “active” and “resting” rates. They did this in various fashions, often with value bars or bins, or pie charts within bins.

Of the 13 students present at the first session, all except two considered Individual Case aspects of the data related to the people in the class and largest or smallest measurements; for example, “Resting... 1. When resting Peter has a high heart rate. 2. When resting nick [s.] callum and bianca have a low heart rate.” Because of the enormous interest in the activity and differences in individual values for the class members, Jenny did not discourage the discussion of these points. Four of the students followed up on this by including Explanatory/Context comments in their text boxes in TinkerPlots, for example, “running 1. nick is the most unfit. 2. krystal is the fittest. But krystal may not have been running as much as nick so her heart may not have been working as hard as nicks.” Context was explicitly mentioned by three students in relation to the fitness of the class. These comments could be considered as embryonic hypotheses but only for the context of the data collected for this class, not for any wider population.

As a Summary statement, average of some sort was mentioned by five students with two noting the most common value, while others used the mean as recorded on their graphs by the software: “running ... the average heart rate is around 140 because the computer has added the heart rates up and divided that by the amo[un]t of people in the class.” Spread was discussed by four students, one providing the following comments: “1. There is a huge difference between highest and lowest on both of the graphs. 2. It makes it different depending on the range of the scale.” The difference in scale for the two graphs (60-80 vs 100-176) was the focus of concern in class discussion, with students again saying it was not easy to make comparisons. Near the end of the lesson students were taught how to make the scales of two graphs the same but few had time to do so.

Four students produced graphs but no accompanying comments in text boxes. Two students used bins to organize the data but did not explain the representation. Those who created scatterplots of “active” and “rest” rates either made no comment or said the graph was not easy to interpret. Only two students presented hat plots for the two data sets with the same scale. These students had no other graphs in their files and made no comments in text boxes. They may have deleted earlier graphical representations and the absence of text indicates the two students may have followed the procedures with little understanding. In Session 1 no students entered comments in text boxes related to Shift, Signal, or Sampling. All students created at least two plots in TinkerPlots and eight were considered to produce Unistructural responses in their text boxes as shown in the quotes in this section. One student might be considered to be providing a Multistructural response in considering three aspects of the data. Four students who entered no comments in text boxes were considered prestructural in their observed learning output. The only move beyond the data was seen in attempts to explain the data within the classroom context. In considering the 104 cells in the analysis matrix for the 13 students and 8 elements of the Beginning Inference Framework, 31% were annotated indicating students had produced graphs or text.

The output shown in Figure 3 is typical of the student interest in the heart-rate values for the class. The plots illustrate the students’ use of bins, scale, and value plots.

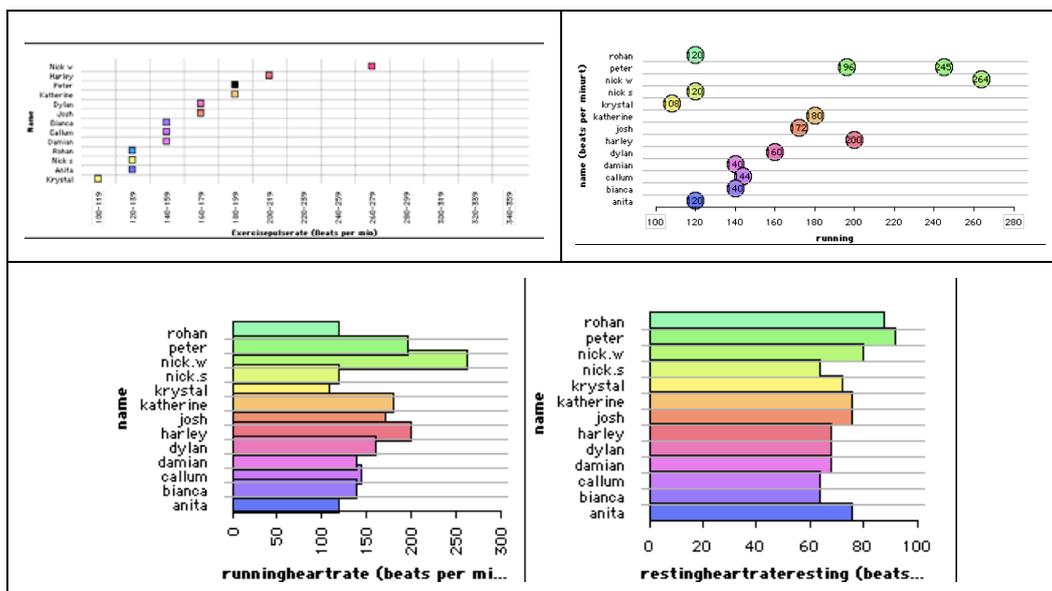


Figure 3. Representations of individual heart rates for each of the students in the class created in Session 1

4.2. SESSION 2 – HEART-RATE DATA

Much of the time in Session 2 was spent by Jenny explaining the concepts related to hat plots and percentage and hence there was little opportunity to experiment with different representations of the heart-rate data sets and compare them using hat plots. Besides hat plots, discussion focused on the ranges of the two data sets, their means and medians, and the necessity to have the same scale for each graph. All students, except one, created hat plots for the two data sets. Student output in TinkerPlots files was collected at this point for analysis. All students created two similarly scaled graphs and included at least one hat plot. Three students included no text comments. The other eight students made comments about the Spread of the data, some of which were related to the central 50% of data covered by the crown of the hat. No one commented on Individual Case values; this may have been because they had studied the data earlier or because of the relatively structured nature of the lesson. Only one student commented on the Explanatory/Context in terms of exercise and two noted Sampling issues, related to the validity of the data recorded. The T-R reported that in overheard conversations a few students were unwilling to suggest a “difference” in heart rates at the two times because there was some overlap in the data. The graphs and text from one student are shown in Figure 4. The student’s response was considered to employ the elements related to Shift in commenting on the different ranges, to Signal in noting and comparing the range of the circles “together” and “all spread out,” and to aspects of Sampling in questioning the validity of some values and wanting some to repeat the trial.

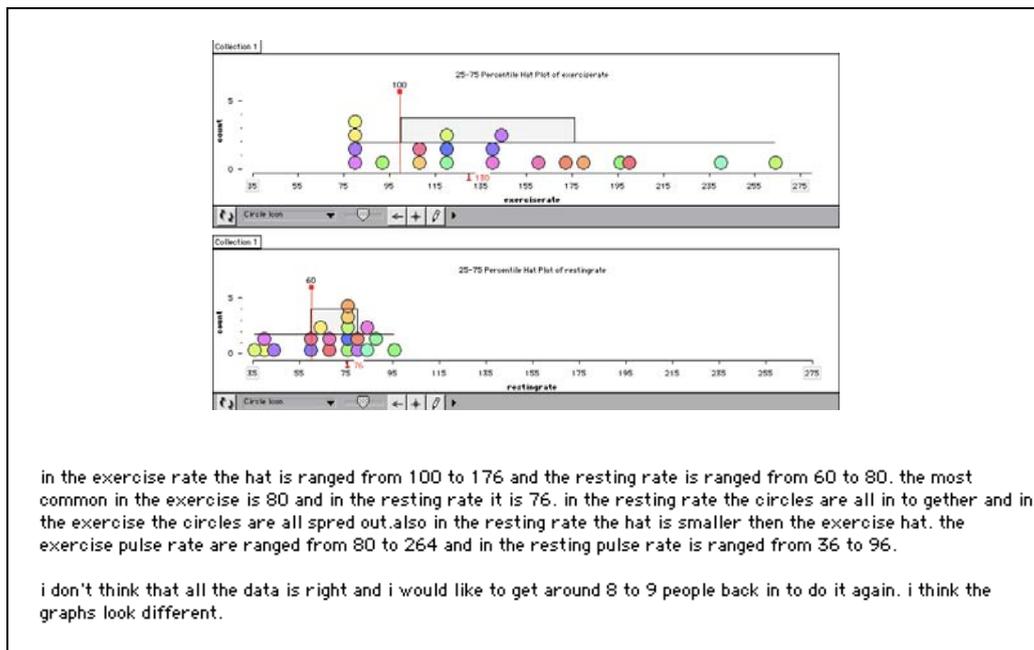


Figure 4. Student extract from Session 2

The differences in the elements of the Beginning Inference Framework addressed in Session 2 compared to Session 1 are quite striking, reflecting the teacher input of the session. Of the observed student outcomes in the TinkerPlot files, Summary of the graph is the main overlapping feature, whereas there was a strong shift from considering individual cases to considering spread and the middle 50% of the data in relation to the

hat plot in Session 2. Only one student commented specifically on an average value (Summary), whereas five noted Shifts in the data sets, three commented specifically on the centre 50% (Signal), and eight discussed Spread. Only one commented on Individual cases, and two made comments on the Sample (as in Figure 4).

In Session 2 three students made no comment in the text boxes, two made single Unistructural responses in relation to spread, and six made two or three comments of a Multistructural nature. Of the 88 cells in the analysis matrix for the 11 students and 8 elements of the Beginning Inference Framework, 43% were annotated indicating either a graph or text had been produced. The purpose of the session on developing the appreciation for hat plots had focused student attention on ways of distinguishing two data sets, a precursor to larger aspects of informal inference.

At the end of Session 2, the T-R led a discussion about the validity of the data collected. Although they had not talked about it before, students could offer suggestions about why the data could be in error. These included miscounting or not concentrating. To remedy the situation students suggested having someone else take the heart rate, checking the method used, or having the same person take all of the measurements.

4.3. SESSION 3 – ARM-SPAN DATA

The T-R returned and helped Jenny teach a session where the class analyzed the data collected on the arm spans of the middle-school students (see Section 3.2). At the beginning of the class, the T-R assisted with recording the class list of terms. It was decided to enter four attributes on the cards: name, grade, arm span, and sex. Students were then asked to think about different ways of exploring the data, with questions like “What does it tell you?”, “Is it meaningful?” They were encouraged to use the terms on the whiteboard as starting points for their investigations. They were asked to keep all graphs and write a text box about what each one told them. The students were actively involved and discussed what they were doing with each other. Most students had created at least three graphs by recess. At recess, students were provided with morning tea and sat around a large table, discussing different ways of looking at data. They summarized aspects of the data, for example the differences between some grades (e.g., 5 and 6, or 6 and 7) but not others (e.g., 7 and 8). This introduced the topic of growth spurts, which led to a discussion of differences between boys and girls. Although the means and medians for boys and girls were nearly the same, the spread for boys was much greater. The tallest and shortest grade 7 boys talked about their advantages and disadvantages. Talk about spread led to making the point that one needs to know more than the mean or median of a data set. Then the issue of sample size was raised because there were different numbers in each grade and of each gender. The most difficult question discussed was related to representativeness, for example when asked, “Which ‘group’ was the data we have collected true for?” and in particular how would they tell if the data were “true” for students in grade 5 to 8 in Tasmania, someone suggested measuring the arm span of every grade 5 to 8 student in Tasmania. The concept of sampling a few schools appeared to be difficult for all of the students.

After recess students went back to the computer lab and were asked to write a Final Report: to choose a comparison of interest, and using the class dictionary on the whiteboard as a guide to write their report. Different tools were to be used, including hat plots, dividers, and percentages, and students were asked to recommend the ones they thought were most useful in making the comparison. The saved files from Session 3 were collected for analysis.

The spread of responses across the Pfannkuch elements was much greater than in the first two sessions. Three of the students created separate files for the Final Reports and most appreciated the task of summarizing their work and providing some sort of evaluation of their use of TinkerPlots. The only aspect not covered was that of Sampling, a topic that the T-R reported discussing at recess and one that she felt the students found difficult to comprehend. Due to the lack of an initial discussion of a wider population and hypothesis, this is not surprising. Three students considered the context of student growth with grade, which may have been a result of the discussion at recess. Four students appeared to address Hypothesis Generation for differences across grades but these were locally-based observations. It was mainly the use of language that separated these from the Explanatory/Context comments. Except for one student who did not write in text boxes, all students discussed Summaries of the data and their plots, including averages and observed differences. The word “shift” was not used with the box plots but change was addressed by six students. Spread, with nine specific comments mainly about range, and the Signal in the central 50% of the data, with six specific comments, were considered at about the same level as for Session 2 but as not all students used hat plots in Session 3, the references were generally encouraging. Irrelevant information not related to arm span but to the number of boys or girls in groups was provided by six students, two of whom had not been present during Sessions 1 and 2. All students made some comment on Individual Cases, presumably related to their encounter with a new data set where they knew the participants.

Although some students struggled with their informal inferences, they appeared to be appreciating the process involved. The following two comments by one student are associated with the graphs displayed in Figure 5. The response was judged to be a Relational response in the context of this study.

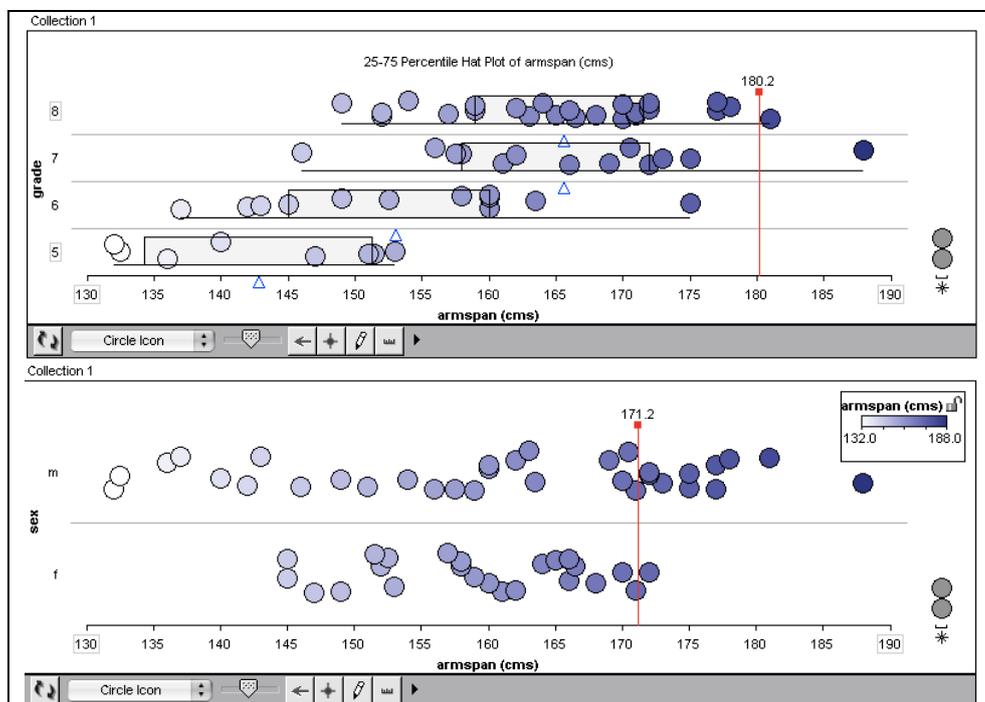


Figure 5. Figures related to student comments for arm span data

[I] com[pared] the arm span and grades. there was an 11 cent[i]meter difference between the mean of grade 5 and 6, and a 12cm difference between the mean of grades 6 and 7, but between grades 7 and 8 there was almost no difference at all in the mean. This tells me that the students do most of their growing in grades 6 and 7. [Figure 5, top]

I compared the armspan with boys and girls. The range of the boys results was much m[o]re spread out[.] They went from 132cms to 188.1. The girls went from 145cms to 172cms. I think this is because girls are usually smaller than boys. [Figure 5, bottom]

Another student used hat plots for the comparison of the grade 5 and grade 8 students. The descriptive account of what the student found out about the data is quite extensive (the student uses “hat” for the crown of the hat). The text box entry relates to Figure 6:

i find the difference between the age and the size of the arm span interesting. i used the hat because it shows me that the hat is covering 50% of the dots. in the grade 8 the hat is around 159.0 to 171.0 and in the grade 5 it[']s around 136.1 and 151.6. i think the hat is easy to use because it shows me where the most armspans are ranged from. i used the mean and the median because it shows me that the average in the mean in the grade 8 is 165.6 and in the grade 5 it is 145.1 but the median in the grade 8 is 164.9 and in the grade 5 its 147.0. in the range of the grade 5 the smallest is conner with 132 and the highest is bryce with 163.5. but in the grade 8 the lowest is kasey with 149 and the highest is tim with 181.

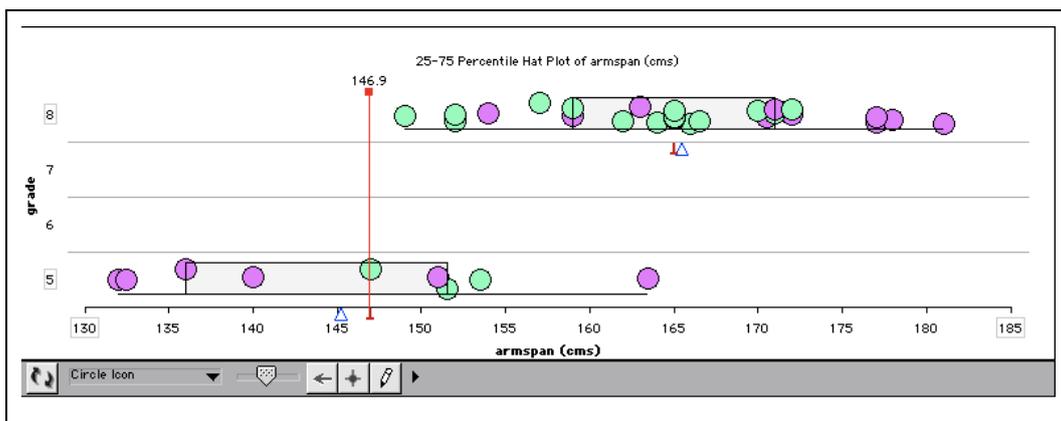


Figure 6. Student plot from Session 3 using hat plots

The student then produced the graph in Figure 7 and explained the difference for grades 5 and 8 in terms of percentages related to a central “average.” This is an interesting comparison but the criterion for determining the central average is not explained: “with the div[i]der and the % i can see that 60% of the grade 8s are the average and 40% is higher th[a]n the average. in the grade 5 44% is the average and 56% is lower th[a]n the average.” Overall this student’s response was judged Relational in the context of this study.

Acknowledging that students during Session 3 were asked to keep all of their output and to prepare a Final Report, it is not surprising that a wider coverage of Beginning Inference Framework elements occurred. All students volunteered that they liked using TinkerPlots generally but some still reported that they found hat plots difficult. It is likely that this difficulty is related to their previous mathematical background, especially their understanding of percentages.

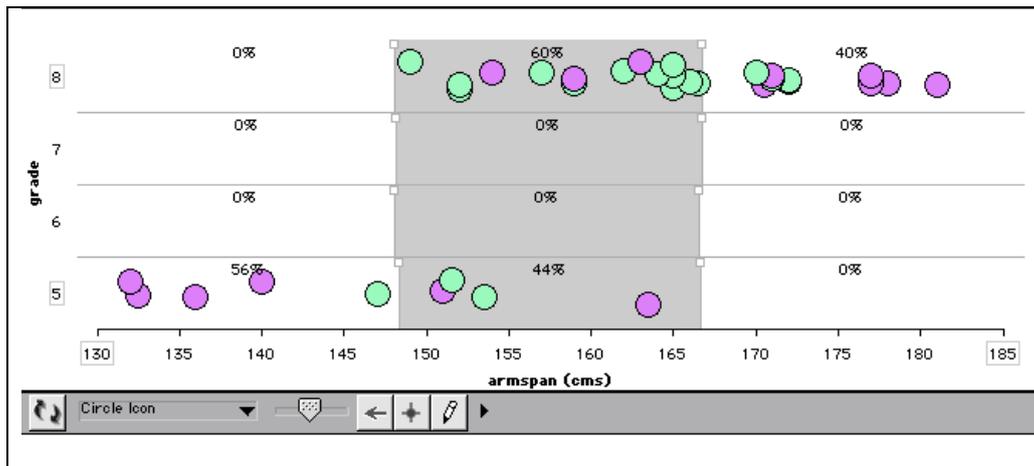


Figure 7. Student plot from Session 3 using dividers

In this session, where students experienced the most freedom in completing their analyses, only the aspect of Sampling did not receive any attention. With a new data set students were again interested in Individual Cases. The consideration of Explanatory/Context was often related to the informal Hypothesis Generation about the data set. The linking of consideration of context/hypotheses and the Summary of plots and Spread appeared to result in Relational responses by six students, whereas six others appeared as Multistructural in sequencing Summaries of the plots, Spread, and Individual Cases. Only one student did not fill in at least some text boxes. Of the 104 matrix cells in the Beginning Inference Framework for the 13 students and 8 elements, 57% indicated that students had produced a graph or written text.

4.4. SESSION 4 – GROWTH DATA

In the session 3½ months later at the start of the following school year, the T-R led a discussion of terminology, including the specific terms hypothesis and evidence. Students contributed to the discussion, which concluded that hypotheses were the questions they wanted to answer and evidence related to how they knew something, for example to support their hypotheses. Students were then given TinkerPlots files to work on.

The specific questions in the text boxes (see Figure 2) encouraged the students to consider the eight components of the Beginning Inference Framework. Two girls answered the questions on print-out rather than using text boxes. A few students, however, continued to struggle, leaving some text boxes unfilled. Whether this was due to lack of understanding, lack of literacy skills, or general unwillingness to complete the task is unknown.

The T-R reported that the conversation among the students was of a higher quality than the written responses, as students were reluctant to compose expressions in written words. The difficulty with literacy skills is apparent in some of the responses. The relationship between the questions about hypotheses and descriptions of the graphs was confusing for some students. As well, meeting a much larger data set for the first time made it difficult for some students to decide if a relatively small difference in appearance constituted a genuine “difference.”

As the graphs in Figure 2 were not altered by the students in this part of Session 4, student responses from the text boxes from three students for the five questions are presented in Table 2.

Table 2. Selected extracts from three students to the questions presented in Session 4

Consider the difference in heights between the males and females at age 2, at age 9 and at age 18. What changes do you see?	How do the graphs help you decide if there is a difference between males and females at the three ages?	What would you hypothesize about the differences?	How do the graphs support your hypothesis?	What questions would you ask about the data?
They are getting taller. At 18 years the boys are about 10 cms taller. [Student 1]	the hat plot shows the difference between boys and girls at the 3 ages.	up till about 11 years boys and girls are a similar height. when get to 18 years the boys get taller than the girls.	the graphs show that the hat plots separate when the boys and girls turn 18.	what is the average height for both boys and girls at 18 years. why do girls reach their heights early before boys
the spread becomes different the older they get but the real growing happens in the 9 years between 9 and 18. [Student 2]	the hat plots are the main help they make it easier to compare the girls and the boys.	boys in general are higher, girls are smaller but when younger in children they appear the near same.	I actually got the hypothesis from the graphs so they support what I have been saying considerably.	maybe do a graph showing the boys/girls at 2 and a bar underneath at 18 so you can see how they have grown in this way you would need names near each dot.
<p>when the females are younger there are more near the end of the line. when they get old they start to go near the start of the line.</p> <p>2 years-the males are spread out more than the girls in the hat plots and the hat plots overlap each other.</p> <p>9 years- there is not much difference between the middle 50% and most of the hat plots overlap each other.</p> <p>18 years-there is a little difference between the middle 50%. the hat plots start to separate from each other. [Student 3]</p>	<p>the hats help me see the middle 50% (crown) and the 1st and last 25% (brims) between the males and the females.</p> <p>when they are stacked it shows me how many people are at that height.</p> <p>it also shows the range of people in the hat.</p>	that the males would be taller than the females when they were younger and older but in the middle age the females would be taller.	<p>it shows that the males are the tallest in the younger and older age and the middle age a female is taller than the males.</p> <p>at 2 years the males are taller than the females.</p> <p>at 9 years the males were the tallest but one of the females was taller than the males.</p> <p>at 18 years the males were the tallest of all.</p>	<p>is this information we have sure or is it guessed?</p> <p>have we got the real measurements or are the measurements rounded off?</p> <p>what is the range of males at the age of 9?</p> <p>what percentage of girls are under the height of 100 cm?</p> <p>what is the average height for a 14 year old?</p> <p>what range can a 14 year old grow to?</p>

The text in Table 2 has not been edited, reflecting exactly what was written in text boxes. Not all responses were as articulate as these. In a couple of cases other students produced similar responses to these indicating that discussion and sharing of ideas took place among class members. Student 1 was considered to have addressed four of the elements in the Beginning Inference Model: Hypothesis Generation (up until 11 about the same then boys getting taller), Summary (at 18, boys about 10 cm taller), Shift (hat plots separate at 18), and Explanatory/Context (why do girls reach heights before boys?). This was judged a Multistructural response as a sequence of comments in these four areas. Student 2 also addressed four elements in the model: Hypothesis Generation (boys in general higher, when young nearly the same), Summary (hat plots easier to compare boys and girls), Spread (different as they get older), Explanatory/Context (real growing from 9 to 18). These comments were considered to be more integrated and hence Relational in nature. Student 3 addressed all of the elements in the model: Hypothesis Generation (males taller when younger and older; females in the middle), Summary (range and heights [frequencies]), Shift (at 18, hats separate), Signal (at 9 years, middles the same), Spread (shows range, hats overlap), Sampling (is information true or guessed?), Explanatory/Context (what range at 14), and Individual Case (at 9, one female taller than all males). This response was judged to be Relational.

Of importance in the responses in Table 2 is the continued use of the article “the,” apparently indicating that the students were considering their hypotheses in relation to the data set presented rather than a larger population of children, which would be referred to without the article. Hence, it appears that although some steps had been taken to use the language of hypotheses across groups, it does not represent a genuine attempt at generalization to a much wider population.

When given the opportunity to change the plots in the file to tell the story of the data set better, 12 of the students made some changes to at least one of the graphs in Figure 2. Of these, five made no additional written comments, for example changing to square or “people” icons and un-stacking the data, sometimes recombining the males and females in the data set, or using a small number of bins. Others commented on the fusing of icons or using of dividers as helping them to see crucial aspects of the data for telling the story. Three boys scaled the three graphs from 80 to 200, one with un-stacked square icons and no hats. One of the other representations is shown in Figure 8. Two of the comments show that the task was appreciated in terms of helping to justify the responses provided earlier in the text boxes.

Student 4: this shows them all at the 18 year old range, this is particu[lar]ly us[e]ful for showing how the people have grown over the time period[.] there proba[b]ly could have been a graph made in the 12 year old range to bridge that huge gap between 9 and 18.

Student 5: the graphs that i have just made makes it [ha]rder to disting[ui]sh the h[e]ights but easier to compare them to their selves 16 [yrs] down the track. this ma[d]e it easier to compare [their] 2yr olds to the 18 yr olds and so forth. but doing this also made the hats harder to distinguish.

Two girls produced scattergraphs with height at age 18 on the horizontal axis and height at age 9 on the vertical axis. One is shown in Figure 9 and is color-keyed by gender. As the students had had virtually no experience with interpreting scattergraphs,

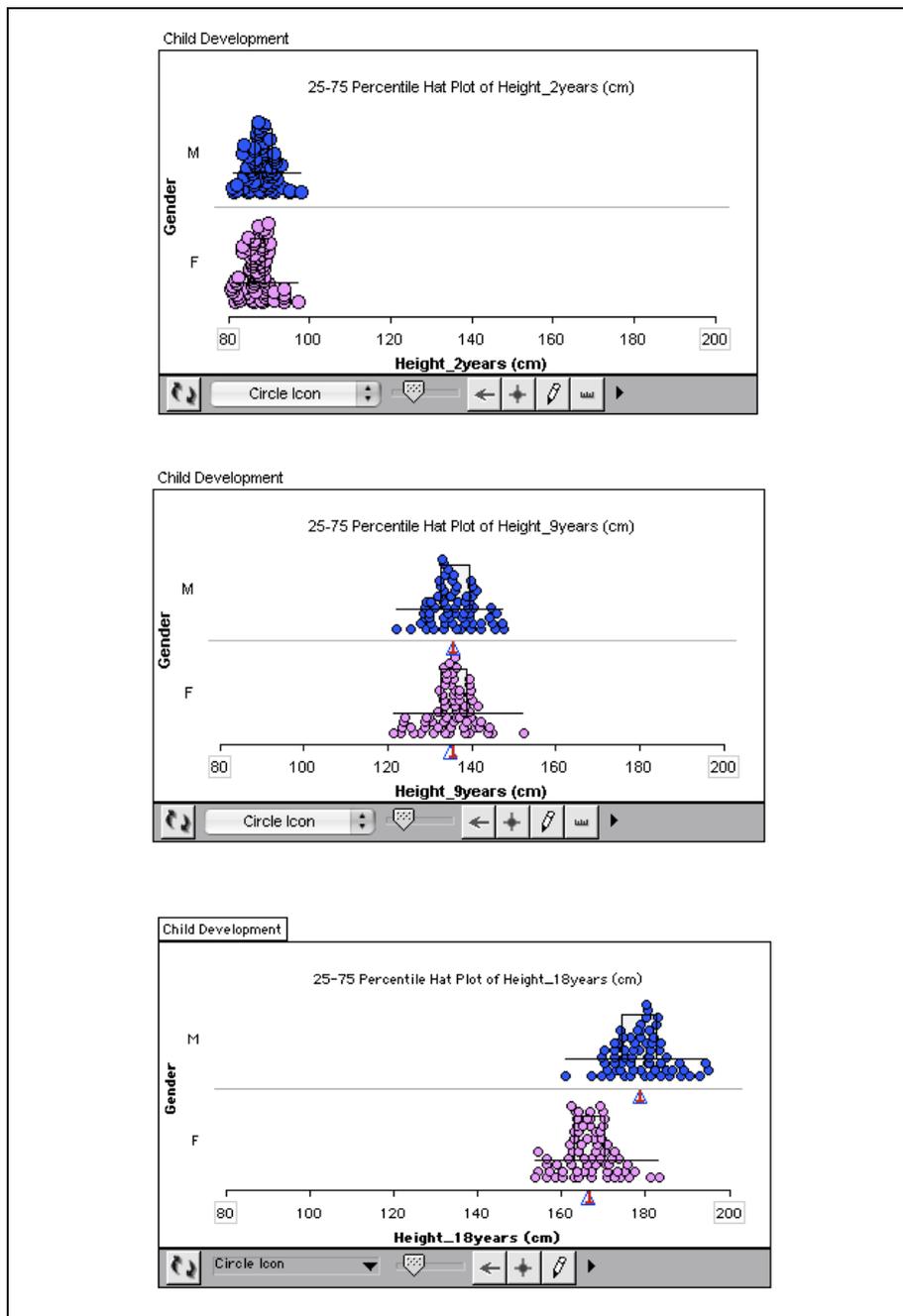


Figure 8. Scaling of height data by students in Session 4

the comments partly missed the point. “This graph shows that as you get older the more you grow. [A]nd that males are mainly taller than females but some are the same.” The attempt, however, demonstrates a growing appreciation of the task of generating hypotheses from the data.

In the final session, the structured questions scaffolded Relational responses in 11 out of 15 cases with the other 4 being Multistructural. The interest lay in how students expressed themselves. Overall these responses showed what the students had picked up from the previous sessions and their beliefs in the usefulness of dividers, reference lines,

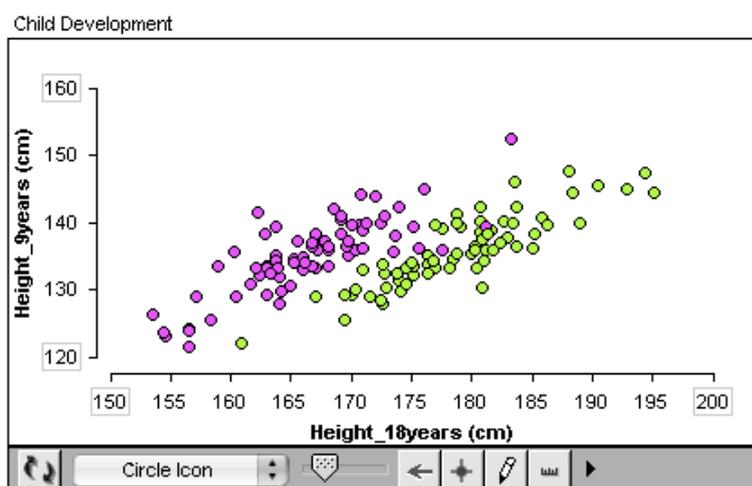


Figure 9. Scatterplot for height data

and scale in telling stories about data sets. Of the 120 matrix cells in the Beginning Inference Framework for the 15 students and 8 elements of the framework, 67% indicated text responses. The fact that text boxes were provided with questions may have supported this increased percentage.

4.5. SUMMARY

The clustering of students' observed outcomes within each session is summarized in the previous four sections. Because of the different increasing levels of facility with TinkerPlots and the different degrees of scaffolding across the teaching sessions, it is somewhat difficult to compare the structural complexity of students' responses written in text boxes over time. The levels reported in Table 3, however, show higher levels of observed responses across sessions. Jenny felt that in terms of student motivation and application to the tasks these observations were indicative of higher achievement than usual for her class. Further, a perusal of the comments for the three students reported in Table 2 shows that appropriate language was being taken up by these students.

Table 3. Summary of structural complexity of observed responses over four sessions

Session	1 ^a	2 ^b	3 ^c	4 ^d
No. of students in class	13	11	13	15
No text response (Prestructural)	4	3	1	0
Uni-structural	8	2	0	0
Multi-structural	1	6	6	4
Relational	0	0	6	11

^aHeart-rate data entry. ^bHeart-rate hat plots. ^cArm-span data entry and analysis. ^dInterpretation of height/age/gender data.

Eight of the students were present for all four sessions where data were collected. A summary of their levels of text box entries is presented in Table 4. The improved observed performance of these students reflected that of the overall group. That these students could take up many of the elements of the Beginning Inference Framework and

use them with guidance represents a first step to informal inference. The comments of Students B and C for Session 4 are among those presented in Table 2.

Table 4. Levels of observed response for eight students who were present for all four teaching sessions

Student	Session 1 ^a	Session 2 ^b	Session 3 ^c	Session 4 ^d
Student A	U	U	M	M
Student B	U	M	R	R
Student C	U	P (no text)	R	R
Student D	U	M	R	M
Student E	U	U	M	R
Student F	M	M	M	R
Student G	P (no text)	M	M	R
Student H	U	P (no text)	M	R

Note. P = Prestructural, U = Uni-structural, M = Multi-structural, R = Relational

^aHeart-rate data entry. ^bHeart-rate hat plots. ^cArm-span data entry and analysis.

^dInterpretation of height/age/gender data.

5. DISCUSSION

5.1. STUDENT CHANGE IN RELATION TO BEGINNING INFERENCE

In relation to the research questions that guided this case study, the first concerns the observed learning outcomes for the students in relation to beginning inference. Using the Beginning Inference Framework adapted from Pfannkuch (2006b), the students used the facilities offered by the TinkerPlots software to address elements associated with Summary, Shift, Signal, Spread, and Individual Cases. At the grade 7 level it was more difficult to distinguish the elements of Hypothesis Generation and Explanatory/Context as context was essential to both elements and to the students' language of questioning and speculating. Sampling was the most difficult element for students to assimilate. It is likely that this difficulty arose from the order in which the lessons proceeded, the limited life experiences of the rural students, and the delayed emphasis on the rather abstract nature of the sample-population relationship. As noted, the T-R found from her interaction with the students that they were having difficulty with it.

In terms of the observed learning outcomes as assessed from the TinkerPlots files created by the students (with two exceptions being hand written for the final session), the SOLO levels of observed outcomes, as well as the percentage of elements addressed, increased with the sessions. The scaffolding of the classroom experiences as well as the nature of the questions in the final session undoubtedly contributed to this increase, but this was indeed the aim. The long-term nature of this understanding past Session 4 is impossible to determine from the study.

According to Jenny, the students had had no prior experience with graphing software and their graphing experience appeared restricted to bar charts, and in a couple of instances pie charts. Students, however, adapted easily to the drag-and-drop features of TinkerPlots and most, after the first session, were able to select appropriate variables to place on the axes. In terms of the goal of comparing two (or more) groups, the majority of students appreciated the value of hat plots. In the early stages they had some difficulty in judging whether "a difference" existed at all, even when there was very little overlap of two data sets, whereas later some students were willing to claim as genuine differences what appeared to be quite small differences. The use of the hats, both created and

interpreted by students, proved helpful to some students in describing shifts in data sets, although the language employed was often quite colloquial (see Table 2). A few students at the end still preferred to put the data in bins but they could explain, for example to the T-R, what the hats represented. Students participated well in the class discussion but did not enjoy writing comments in the text boxes in TinkerPlots to record their reasoning. The written comments were not of the same caliber as the notes recorded by the T-R during the sessions. In two instances in the final session girls wrote their thoughts on a hard copy of the TinkerPlots output rather than using the text boxes. The hand-written text was much more extensive than anything presented in a text box in the earlier sessions by these two students. This and the observation of the literacy and typing levels of students lead to the suggestion that a second order of difficulty is introduced for many students when asked to express ideas not only in words (for example on paper) but also in typed text in a text box. This leads to the conclusion that the levels of response observed are likely to be minimum levels for these students.

The observed interaction of the learning framework, the teacher's preparation, and the software is important in evaluating the students' observed outcomes and suggesting changes. The teacher and the T-R were introduced formally to the Beginning Inference Framework after Session 1, when it was decided by Jenny to continue working with her students. The T-R implemented the aspects in her planning and at various points she raised them with the students. At no point, however, were students provided with notes or handouts suggesting they consider each component of the model. Due to the limited number of teaching sessions, not a great degree of reinforcement was possible. The T-R expressed some surprise, however, at what the students remembered over the summer period (3½ months) between Sessions 3 and 4. It is clear from the analysis that in a structured setting the students could apply many of the components of the model and some could relate them together in meaningful ways. That the students had used the TinkerPlots representations to assist in making their comments is evident from the explicit phrases in the text boxes. Providing a simpler version of a box plot, accessible to these students, appeared to be of assistance to those who made Relational comments. As noted throughout the study, however, some students employed dividers, reference points, means, and medians to supplement or replace hat plots in reaching decisions about the data sets.

That the point reached by the students in this case study along an imaginary "informal inference" continuum was not very far past "beginning informal inference" is clear. They had, however, begun to address some of the criteria set by Rubin et al. (2006). The progress observed by the T-R, herself an experienced teacher, however, was considered extraordinary given her observation of students' starting points and the general background of schools in the larger project. The students certainly made progress and all except one claimed in a feedback form that they would use TinkerPlots again.

5.2. IMPLICATIONS FOR FUTURE RESEARCH AND IMPLEMENTATION

The second research question follows from the first in the light of the experiences during the four classroom sessions. The limitations of this case study are those of most educational research in less-than-ideal settings. The students were from a rural setting where formal education is not always valued; as is seen in some of the student extracts, literacy levels were low for grade 7. Eighteen different students were present in Jenny's class for at least one of the four sessions, whereas only eight were present for all four. Three areas for changes in future research and implementation are considered.

The use of the Beginning Inference Framework adapted from Pfannkuch (2006b) was very useful for most of the eight elements. The interaction of the elements Hypothesis Generation and Explanatory/Context, however, caused some difficulty in allocating students' observed responses to one category or the other. In explaining the context as they saw it, students sometimes used it as a foundation to suggest a hypothesis. In the future, increased emphasis on sampling and the relationship of samples to a population may assist students in being more explicit in generating hypotheses more generally than for the specific context, for example, of their class. The issues associated with using the framework for planning teaching and for assessment lead back to the general question of how many and which of the elements can or should be introduced over time during the middle years. For teachers who themselves have little experience with inference, there appear to be difficulties in absorbing the content to the extent of appreciating what their students experience as learners and hence in developing pedagogies that are appropriate. A large-scale study considering alternative orders of introduction of the elements might offer guidance in this area.

Although Jenny had experienced professional development before the teaching sessions, had assisted other teachers, had talked to the researcher and T-R, and was enthusiastic about introducing the software to her class, at times she lacked confidence in relation to the goals of an investigation in terms of beginning inference and turned to the T-R for guidance. In these situations the T-R was fairly directive in intervening with the class. Hence, it is likely that a more intensive initial introduction to the Beginning Inference Framework needs to precede a teacher's planning and implementation of similar activities. Teachers' enthusiasm for the software must be complemented continually with the goals of the middle school statistics curriculum. Whether a framework as complex as the one used in this study should be explicitly introduced to grade 7 students at the start of a unit of work is debatable. Further research would be needed to decide and the author has concern that it would be overwhelming to most grade 7 students, certainly those similar to the ones in this study.

The issue of literacy levels and some students' reluctance to complete text boxes in TinkerPlots suggests that future research could record student conversations while involved in the data analysis or that after the final session individual interviews could take place with the students involved. Although it may be assumed that in this computer age students are able to readily type into text boxes, this was not the case for most of this class. Future intervention with grade 7 students might allow a choice of how the conclusions of investigations are recorded. It is possible that students could print out their graphs and handwrite their reports and inferences on them. The requirement to assess observed learning outcomes can be met in several ways that may allow students the best chance of displaying their learning.

5.3. CONCLUSION

Using both TinkerPlots and the adapted Pfannkuch model (see Table 1) appears an appropriate and valuable resource for introducing beginning inference to middle school students. The software and model reinforce each other, explicitly in terms of the elements of Hypothesis Generation, Summary, Shift, Signal, Spread, and Individual Case. The other two elements of the model, Sampling and Explanatory/Context, although perhaps not directly linked to the software, provide cues to considering the other elements in relation to plots, for example to question sample size or make sense of alternative explanations. In terms of assessment, combining the eight elements of the Beginning Inference Framework with a developmental model from cognitive psychology appears to

allow the documentation of students' observed outcomes. As such, it allows the combination of two important criteria: statistical appropriateness and structural complexity (Watson, 2006). Given the observation of students' initial difficulties in judging whether differences in two data sets were small or large, it would appear that the ease of creating representations and using TinkerPlots may contribute to developing intuitions about what might be considered "significant" or "meaningful" differences.

This study presents the hat plot as a viable alternative to the box plot, hopefully satisfying Bakker et al. (2005) in their request to postpone the latter's introduction. In the meantime, the introduction of hat plots appears to encourage the description of the shape of data as noted by Konold et al. (2002) and the comparison of multiple data sets without the necessity of means (e.g., Watson & Moritz, 1999). The progression toward beginning inference, although slow, appears well documented by the Beginning Inference Framework.

ACKNOWLEDGEMENTS

This study was funded by an Australian Research Council grant, No. LP0560543. The T-R for the study was Karen Wilson. Thanks go to Karen for vetting all descriptions in this report and providing a validity check on them.

REFERENCES

- Bakker, A. (2004). *Design research in statistics education: On symbolizing and computer tools*. Utrecht: CD-β Press, Center for Science and Mathematics Education.
- Bakker, A., Biehler, R., & Konold, C. (2005). Should young students learn about box plots? In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education: International Association for Statistical Education 2004 Roundtable* (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/rt04/4.2_Bakker_etal.pdf]
- Ben-Zvi, D., Gil, E., & Apel, N. (2007, August). What is hidden beyond the data? Helping young students to reason and argue about some wider universe. Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), University of Warwick, UK.
- Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.
- Cobb, P. (1999). Individual and collective mathematical development: The case of statistical data analysis. *Mathematical Thinking and Learning*, 1, 5-43.
- Cobb, P., Confrey, J., deSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1), 9-13.
- Fitzallen, N. (2007). Evaluating data analysis software: The case of TinkerPlots. *Australian Primary Mathematics Classroom*, 12(1), 23-28.
- Friel, S. N., O'Connor, W., & Mamer, J. D. (2006). More than "Meanmedianmode" and a bar graph: What's needed to have a statistical conversation? In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth NCTM yearbook* (pp. 117-137). Reston, VA: National Council of Teachers of Mathematics.
- Kirkpatrick, E. M. (1983). *Chambers 20th Century Dictionary* (New Ed.). Edinburgh: W & R Chambers Ltd.
- Konold, C. (2007). Designing a Data Analysis Tool for Learners. In M. Lovett & P. Shah (Eds.), *Thinking with data: The 33rd Annual Carnegie Symposium on Cognition* (pp. 267-291). Hillside, NJ: Lawrence Erlbaum Associates.

- Konold, C., & Miller, C. D. (2005). *TinkerPlots: Dynamic data exploration*. [Computer software] Emeryville, CA: Key Curriculum Press.
- Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In B. Phillips (Ed.), *Developing a statistically literate society: Proceedings of the Sixth International Conference on Teaching Statistics*, Cape Town, South Africa. [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/1/8b2_kono.pdf]
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.
- Moore, D. S. (1991). *Statistics: Concepts and controversies* (3rd ed.). New York: W. H. Freeman.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- Pegg, J. E. (2002). Assessment in mathematics: A developmental approach. In J. M. Royer (Ed.), *Mathematical cognition* (pp. 227-259). Greenwich, CT: Information Age Publishing.
- Pfannkuch, M. (2005). Probability and statistical inference: How can teachers enable learners to make the connection? In G. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 267-294). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Pfannkuch, M. (2006a). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]
- Pfannkuch, M. (2006b). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, 5(2), 27-45.
[Online: [http://www.stat.auckland.ac.nz/~iase/serj/SERJ5\(2\)_Pfannkuch.pdf](http://www.stat.auckland.ac.nz/~iase/serj/SERJ5(2)_Pfannkuch.pdf)]
- Rubin, A., & Hammerman, J. K. (2006). Understanding data through new software representations. In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance: Sixty-eighth NCTM yearbook* (pp. 241-256). Reston, VA: National Council of Teachers of Mathematics.
- Rubin, A., Hammerman, J. K. L., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D3_RUBI.pdf]
- The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5-8.
- Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Mahwah, NJ: Lawrence Erlbaum.
- Watson, J. M., & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, 37, 145-168.

JANE M. WATSON

Faculty of Education, University of Tasmania, Private Bag 66
Hobart, Tasmania 7001
Australia