1. INTRODUCTION

Many issues surround the introduction of technology for learning of mathematics. Clements (2000) provides a summary and a rationale for moving beyond mundane exercises to higher order learning experiences. This view is shared by Shaffer and Kaput (1999) in suggesting that technology offers the potential to find new ways to learn mathematics. The statistical software that is the focus of this report, *TinkerPlots* (Konold & Miller, 2005), provides an example of this objective for technology in the context of the middle school classroom. Suggestions of context-linked investigations to enhance beginning inference through explorations with *TinkerPlots* using real data are appearing in the professional literature (Watson, 2008; Watson & Wright, 2008). Classroom research based on the innovations introduced within *TinkerPlots* has further illustrated new ways to learn statistics. Konold, Harradine, and Kazak (2007) and Konold and Lehrer (2008), for example, describe learning experiences focussed on understanding distributions by building distributions with a random sampler, rather than observing properties of established distributions. Watson, Fitzallen, Wilson, and Creed (2008) report the value of the hat plot (*TinkerPlots*' simplified version of a box plot) to assist students in appreciating the middle and spread of distributions and in comparing different data sets. The reported research from these and other studies involving technology and statistics learning (e.g., Bakker, 2004; Cobb, McClain, & Gravemeijer, 2003) has been based mainly in planned classroom-type learning experiences rather than data collected via in-depth individual student interviews.

This report moves beyond research associated with particular classroom applications of the *TinkerPlots* software for student learning to its application as a research tool to be used in individual student interviews to assess more generally student understanding of statistical concepts. In gauging the success of their software innovations with college students, Chance, delMas, and Garfield (2004) and delMas and Liu (2005) employed interviews centred on their package of computer simulations of distributions. These interviews were specifically aimed at evaluating the software in relation to course objectives, for example in understanding the standard deviation. Konold et al. (2007) also interviewed students in pairs to document the emerging use of the software to assist understanding of distribution following the classroom experiences. There appears to have been no research, however, comparing software and non-software settings for exploring students' statistical understanding in a more general sense, not associated with a particular classroom intervention.

In moving beyond the classroom learning context to collecting data on students' in-depth conceptual understanding, interviewing students individually is generally accepted as the best approach (Burns, 2000, pp. 582-3). It is possible to ask in what ways the *TinkerPlots* software assists students to display their understanding and explore new challenges in an interview setting. It is also possible to ask whether more is learned about students' conceptual understanding with or without the software. In setting up a comparison that would judge the software's ability to facilitate students' exploration of tasks, it is necessary to employ tasks that can be used meaningfully both with and without the technology.

This study tackles the issue of technology as a research tool by exploring the use of *TinkerPlots* software when interviewing school students about their approaches to beginning inference. The

study is set in the background of the literature on "affordances." This term, coined by Gibson (1977) for didactic objects, generally relates to the uses and usefulness of the object perceived by the user. Chick (2007) for example considered the affordances of particular examples in the mathematics classroom to illustrate a desired concept. In the study reported here the object is *TinkerPlots* and following Anne Watson (2003), there is interest in the potential affordances of using *TinkerPlots* to document student understanding of beginning inference by exploring what *is* possible and what *could be* possible. These possibilities relate to the expectations of researchers, the opportunities for students, the flow on to the classroom, and potential efficiencies.

The two protocols that are used in the study have been used in previous studies of student understanding but without the introduction of *TinkerPlots* or any other software (Watson & Moritz, 1999; Chick & Watson, 2001). These paper-based protocols and the basic features of *TinkerPlots* are described in the next section. Three samples of students are referred to in this study. Samples A and B are described in the next section. These two samples are taken from previous studies and the students involved used the paper-based version of the two protocols. Sample C is described in the Method section. These students used the *TinkerPlots* versions of the two protocols. The Method section also includes a description of the background of the students from all three samples – this is important as it was not possible to have the same students address the tasks under both conditions, paper-based and using *TinkerPlots*. Adaptations of the protocols for *TinkerPlots*, the procedures followed and the analysis of data sources are also found in the Method section.

2. BACKGROUND

2.1 The Technology

TinkerPlots is a dynamic graphing software package created for middle school students from a constructivist perspective (Konold, 2007). Data entry takes place either directly using data cards or a table, or by importing files from spreadsheets or web sites. Instead of the user being required to pre-specify the type of representation for a data set, as for example in Excel, data are first presented in a random fashion in two dimensions on the screen (see Figure 1a). Students can then choose which variables, called "attributes," they wish to drag-and-drop to the plot. Dragging the icons in the plot left or right (up or down) creates more or fewer bins, with the possibility of a continuous scale for numerical attributes (see Figure 1(b) and (c)). Tools available for interpreting data presented in plots include reference lines, dividers, hat plots and box plots, group size (n) and percent, median and mean, and labels. Fitzallen (2007) evaluated *TinkerPlots* in relation to its value as educational software and found it satisfied six essential criteria: (i) being accessible and easy to use, (ii) assisting recall of knowledge and representation in multiple forms, (iii) facilitating transfer between mathematical and natural language, (iv) providing extended memory when organising or reorganising data, (v) allowing multiple entry points for abstraction of concepts, and (vi) providing visual representations for both interpretation and expression (p. 24). The use of *TinkerPlots* as a basis for data exploration for students as young as 8 years old (Paparistodemou & Meletiou-Mavrotheris, 2008) has been reported, as has work with students in grade 6 (Ben-Zvi, Gil, & Apel, 2007), and grade 7

(Watson, 2008; Watson & Donne, 2008). The specific value of the hat plot for representational purposes has also been explored (Watson et al., 2008). The hat plots shown in Figure 2, for example, cover approximately the middle 50% of the data in the crowns and the low 25% and high 25% in the brims for height data for samples of adult males and females. Generally those who have trialled *TinkerPlots*, both teachers and students, have found it an engaging and easy-to-use tool for data analysis.

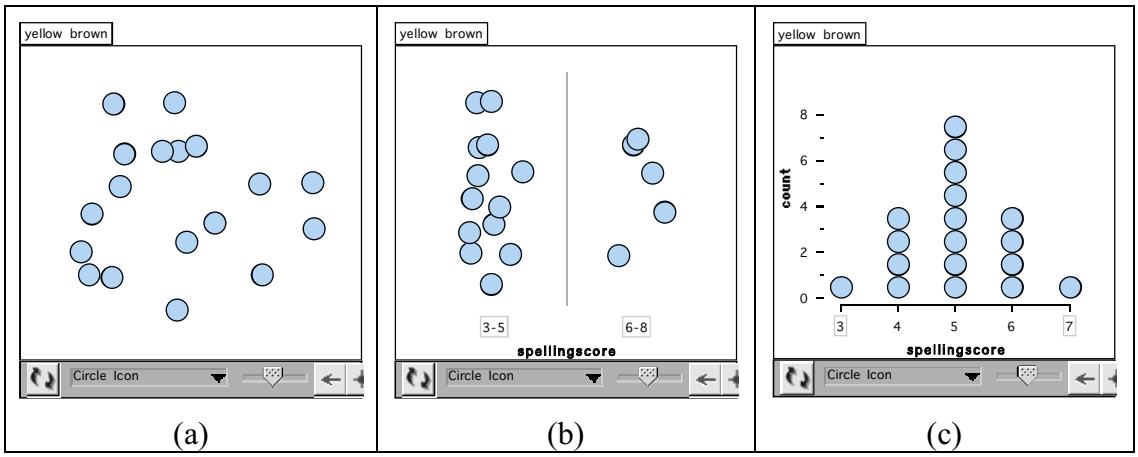


Figure 1: Data floating in *TinkerPlots* (a); Steps in creating scaled graphs for an attribute (b and c).

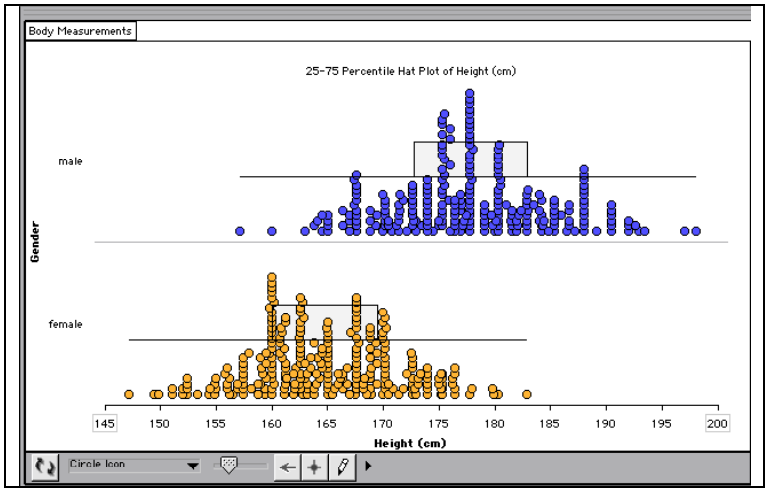


Figure 2: Example of hat plots for comparing two data sets.

2.2 Beginning Inference

There is much debate in the statistics education research community about the use of the term "inference" in data handling contexts when accompanied by adjectives such as "informal" or "beginning." It is not the intention of this paper to contribute to the debate. The assumptions of this paper are based on observations of students aged about 11 to 14 years in the middle grades of schooling. Two aspects of beginning to draw inferences are usually introduced to students. One is the sample-population relationship, with simplified issues of sample size and selection bias. The second is the procedural aspect, which for students of this age usually involves the informal comparison of two data sets or looking at the association of two (measurement) variables. These two applications are usually chosen because they are the most motivating to students at this age and decisions can be based on visual interpretation of graphs (cf. Figure 2). At this age it is difficult for students to absorb and consolidate the interconnection of the sample-population and more procedural aspects of drawing inferences. For a particular data set, they are more likely to become very interested in the difference or association seen within that data set than to remember about the implications for an overall population (Watson & Donne, 2008). Although the importance of both aspects of beginning inference underlie this study of student understanding, it is acknowledged that for these students the "beginning" part of beginning inference is mainly focused on the procedural aspects of interpreting visual presentations and becoming excited about what they can "see."

2.3 The Comparing Groups Protocol

In a study of procedures associated with beginning inference, Watson and Moritz (1999) presented students with four sets of two graphs depicting the spelling scores (number correct out of 9) of students in four pairs of classes: Red and Blue, Green and Purple, Yellow and Brown, and Pink and Black. For each pair the students were asked to determine which class had performed better. The graphs as presented to the students are shown in Figure 3. The progression in the difficulty of comparing the two graphs is seen in the visual presentation of the graphs. Eighty-eight students in grades 3 to 9 completed the task, with little difficulty on the first two parts. For the Yellow and Brown classes, some claimed equality because of symmetry, equal totals, or equal means, whereas others claimed the Yellow class was better because there were more 5s or the Brown class was better because it had a 7. Those using totals often did not recognise the difficulty of unequal-sized groups when comparing the Pink and Black classes. Students had access to a calculator if they chose to use it. Students who claimed Pink was better based on its higher total score were prompted with the question, "Does it make any difference that the Pink class has more students?" This question assisted some but not all students in changing their decisions. Forty-two of the original students were interviewed again with the same protocol three or four years later (Watson, 2001) and of the students who could improve their performance, 65% did so. The students in the study initiated by Watson and Moritz are referred to as Sample A.

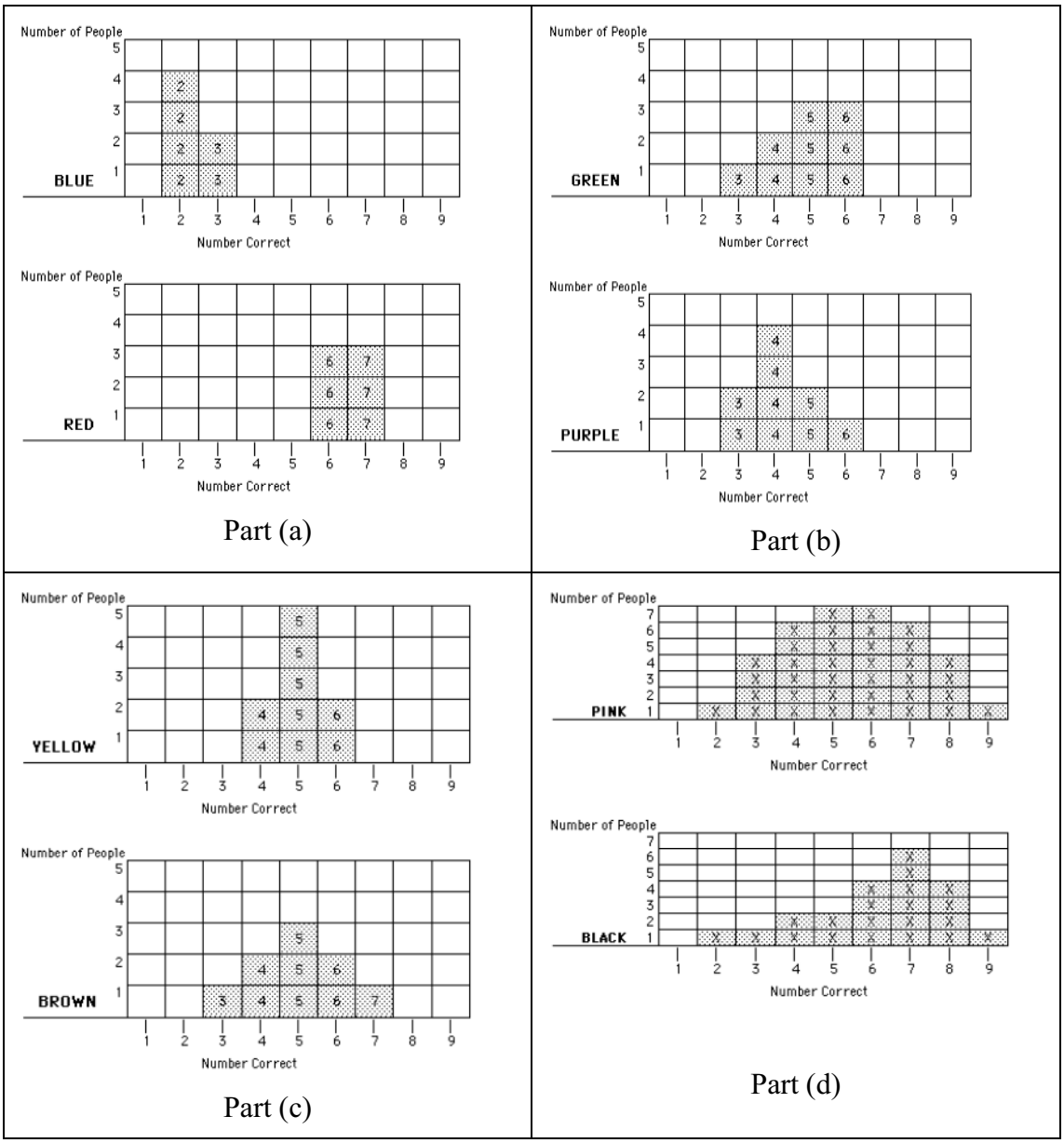


Figure 3: Graphs used for the four parts of the Comparing Groups interview protocol.

2.4 The Data Cards Protocol

The Data Cards protocol was originally devised for individual interviews of students at the same time as the Comparing Groups protocol. Sixteen data cards were prepared with the name, age, eye colour, favourite activity, weight, and number of fast food meals eaten per week, recorded as shown in Figure 4. The complete data set is presented in Appendix A. Gender was not included explicitly as a variable on the data cards but could be identified from the names. After an introductory discussion of students' expectations with respect to the variables, they were asked to explore the cards, devise hypotheses, and provide supporting evidence. Following this and a pilot classroom study (Watson, Collis, Callingham, & Moritz, 1995), an in-depth study in a grade 5/6 classroom was conducted where students worked in teacher-selected, mixed-gender, mixed-grade groups of three on the same task over three lessons. In a fourth session the students presented posters to their teacher, principal, and the researchers. The data from these sessions were analysed from the standpoint of cognitive outcomes (Chick & Watson, 2001), of student perceptions of the collaborative activity (Watson & Chick, 2001a), and of who provides help in collaborative classroom contexts (Watson & Chick, 2001b). The aims of the intervention included the two aspects of informal inference: issues of the relationship of a sample to a population and procedures for exploring differences and associations. The students in the study of Chick and Watson are referred to as Sample B.

Name: Jennifer Rado
Age: 9
Favourite activity: Board games
Eye colour: Green
Weight (kg): 33
Fast food meals per week: 4

Figure 4: Example of a data card.

3. RESEARCH QUESTIONS

Of interest in comparing the affordances of the paper (Samples A and B) and *TinkerPlots* (Sample C) versions of the two protocols is the relationship of the presentation format and the responses given. Acknowledging that there is an interaction between the format and the pre-existing understanding brought by the student to the questions, it is not the hierarchical level of the individual response that is important but the usage of the format across the range of responses in each setting. What affordances are provided for students to display their reasoning using the *TinkerPlots* versions of the protocols that were not available with the paper versions? Are there different understandings observed in the two formats? Are there affordances for *TinkerPlots* in terms of time and efficiency? Are there disadvantages with the use of *TinkerPlots* compared to the paper format? Could these be ameliorated in some way?

4. METHOD

4.1 Adaptation of the Protocols for *TinkerPlots*

The data displayed in the eight graphs in Figure 3 were entered into data cards in four *TinkerPlots* files with attributes: student, spelling score, and class. Each student had a case number; spelling score was the number correct; and class was defined by colour (e.g., Red). A window with the data randomly arranged was shown with the data cards on the screen. A text box contained the same information and questions as the paper version of the protocol. Figure 5 shows the opening *TinkerPlots* screen for the first part of the protocol.

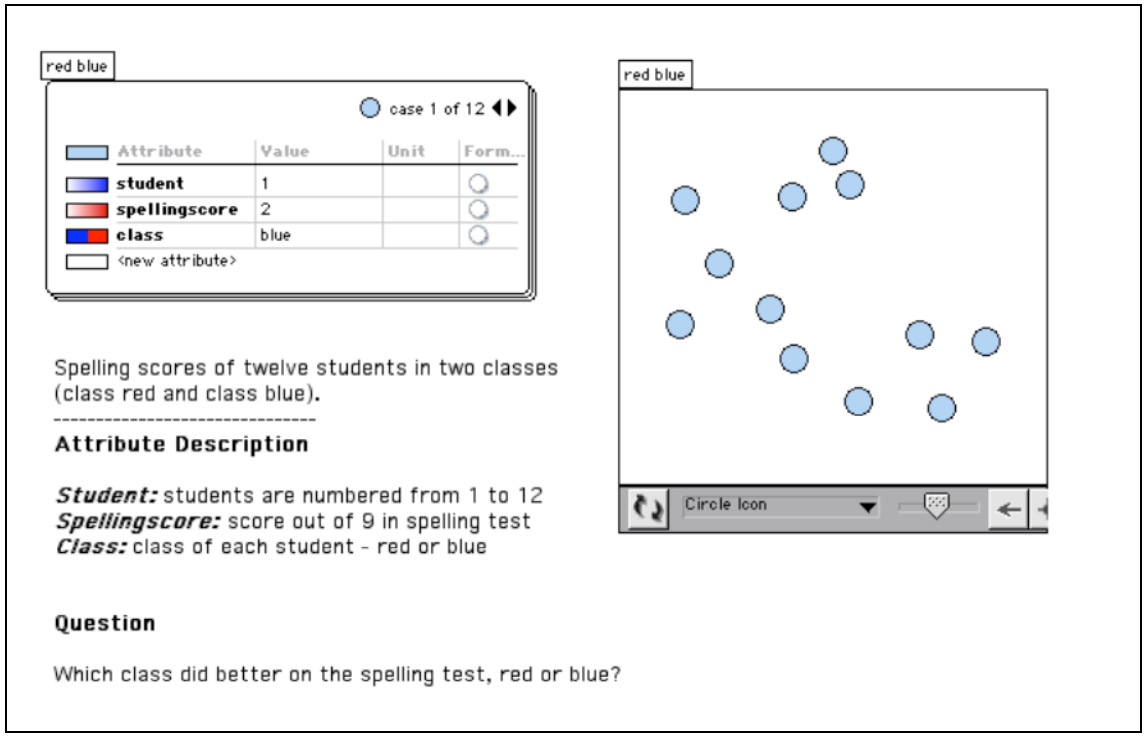


Figure 5: *TinkerPlots* file for comparing the Red and Blue classes on spelling scores.

The major difference for the student experience of the Comparing Groups protocol was in the initial presentation of the data. The students in Sample A were given a paper representation with no opportunity to change it. Students hence needed to be familiar with the format seen in Figure 3. Very few students were confused by the presentation; such students were helped to interpret the first pair of graphs and did not experience further problems with the actual graphical presentation. For Sample C, the *TinkerPlots* group, no initial graph was presented and hence students had to decide the form of representation they desired, or if any was required. The presentation of information was hence quite different in the two cases, even though the data were identical.

For the Data Cards protocol, 16 data cards were prepared in *TinkerPlots* with the attributes in the same order as in Figure 4. These were also printed out and laminated in case students asked to manipulate the actual cards. In fact no students asked to use the physical cards. Again a random

plot was provided and three questions presented in a text box: What interesting things can you find out about this data set? Can you make any hypotheses about the data? What evidence can you produce to support your hypotheses? The opening screen in *TinkerPlots* is shown in Figure 6.

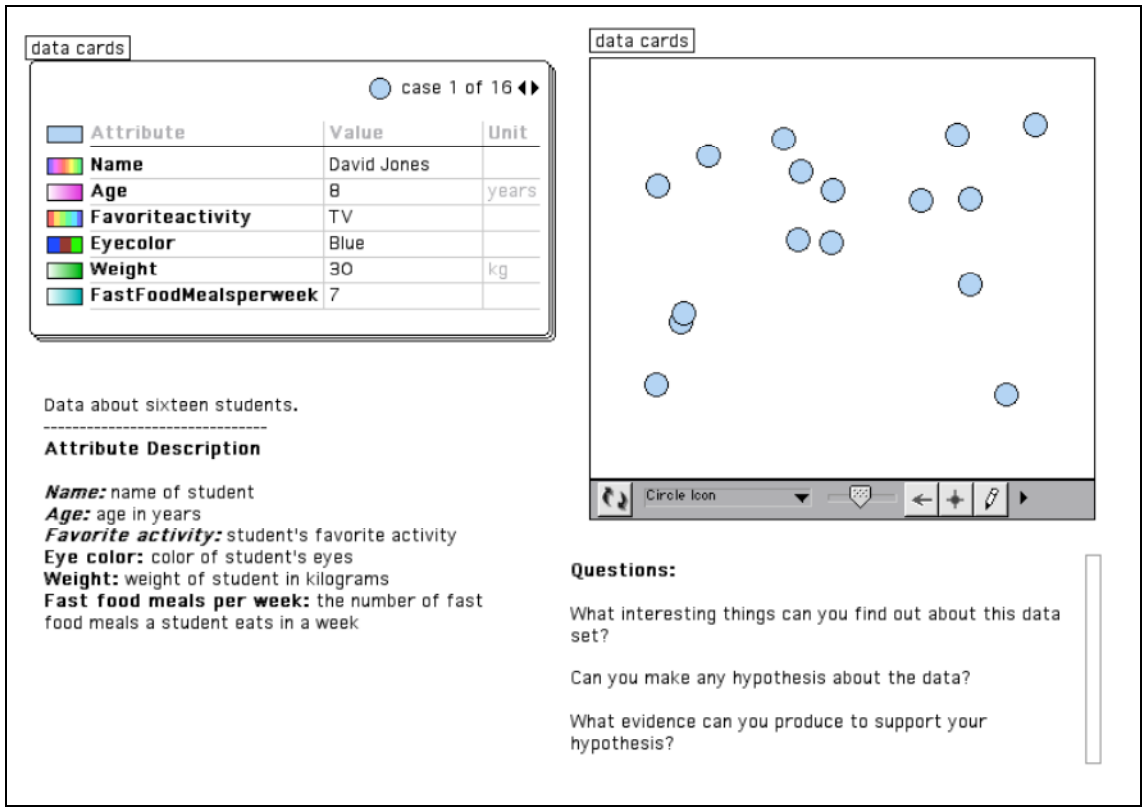


Figure 6: *TinkerPlots* file for the Data Cards protocol.

The initial student experience of the Data Cards protocol was much more similar for students with and without software than was that for the Comparing Groups protocol. The data cards in *TinkerPlots* mimicked very closely the physical data cards used with the students working collaboratively in their classroom. The expectations in each case were the same: to form hypotheses and create representations to support them. Each group had a repertoire of graph types from their previous experiences. For the students in Sample B working in groups these included pictographs, pie graphs, bar graphs, line graphs, and scattergraphs, which were sometimes adapted idiosyncratically to meet the needs of the data being used. For Sample C, the *TinkerPlots* group, the graphing forms included the plots available in the software. Students had been introduced to "bins," stacked dot plots, scatterplots, and circle graphs, as well as hat plots, reference lines, and dividers. Grade 7 students in Sample C had used means and medians.

4.2 Background of Students

For the paper-based Comparing Groups protocol, Sample A, interviewed students were from two Australian states and were in grades 3, 5, 6, 7, and 9. There was no research teaching intervention with the students before their interviews but they had completed a Chance and Data survey as part of the research project. Both states had mathematics curricula reflecting *A National Statement on Mathematics for Australian Schools* (Australian Education Council,

1991), which included Chance and Data as one of five content strands. It is not known, however, the extent to which the curriculum had been implemented in the individual classes from which the interviewed students were selected. The paper-based Data Cards protocol, Sample B, consisted of 27 grade 5/6 students from a single classroom working in groups of three. The same curriculum context existed for these students as for those completing the Comparing Groups protocol. It was known that the grade 5/6 students had worked collaboratively on problem solving in other parts of the mathematics curriculum.

The students interviewed using the *TinkerPlots* versions of the two protocols, referred to as Sample C, were from two schools different from the earlier schools. There were 12 grade 5/6 students and 12 grade 7 students. The grade 5/6 students had taken part in classroom activities over four weeks with *TinkerPlots*, exploring measurement data such as height, foot length, and height of belly button from the floor. The grade 7 students had taken part in four lessons separated by a week each, where they explored reaction time using an applet provided by the Australian Bureau of Statistics *CensusAtSchool* web site again with *TinkerPlots*. All students in Sample C looked at data from their own classes as well as random samples from the *CensusAtSchool* site. These students hence had more direct experience with statistical ideas documented before the interviews than students in Sample A and B and a reasonable degree of expertise with *TinkerPlots* itself. The observed levels of understanding of the grade 7 students in Sample C in relation to their classroom experiences and interviews are reported elsewhere (Watson & Donne, 2008).

4.3 Interview Procedure

Students in Samples A and C were removed from class and interviewed for between 30 and 45 minutes in a quiet room. Students in Sample A were asked other protocols besides Comparing Groups, whereas Sample C students responded to one other protocol, always after the two discussed in this report. Written work, posters, and *TinkerPlots* files were saved. For *TinkerPlots*, when more than one set of attributes was considered, care was taken to take pictures of plots before the attributes were changed, or new plots were brought down from the menu. For all three samples, interviews or collaborative sessions were videotaped and later transcribed or summarised (see Watson & Moritz (1999) for details of initial Comparing Groups interviews).

For the Comparing Groups protocol, students in Sample C interviewed with *TinkerPlots* who did not notice the difference in size of the Pink and Black classes had this brought to their attention with a similar question to Sample A students about whether this would have made a difference to which class had done better. For the Data Cards protocol, using *TinkerPlots*, the grade 5/6 students in Sample C were interviewed first and did not consider defining a gender variable based on the names on the cards. Hence before the grade 7 students were interviewed with the Data Cards protocol it was decided that if they had not considered gender by defining an attribute from their cultural understanding of the names provided, they would be prompted with the questions, “Do you think looking at boys and girls would be interesting? … How could you do that?”

4.4 Data Sources and Analysis

Data for Sample A for the Comparing Groups paper-based protocol are extracted from the examples and summaries provided in Watson and Moritz (1999) and Watson (2001). Similarly the description of all graphs produced by students in Sample B working in groups for the Data Cards protocol is extracted from Chick and Watson (2001). The data collected from the students

who worked in groups of three in their classroom were analysed from two perspectives, that of the level of interpretation observed in relation to the data set and the variables defined by the data cards, and that of the level of representation as observed in the posters created for their final presentations. Because students worked in groups of three, all students were interviewed individually at the end of the intervention to assess their contributions to and understanding of the outcomes produced. The graphs presented in the current paper were used in part by Chick and Watson to determine levels of representation displayed by the 27 students in the class.

For the students in Sample C interviewed in the *TinkerPlots* format, video and transcripts were summarised for each student and categorised by themes identified by the authors on successive readings. For the first two parts of the Comparing Groups protocol, the categories recognised whether students (i) put the data into four bins, (ii) separated the data completely along an axis, and (iii) responded appropriately with the Red or Green class having performed better. For the third part of the protocol three more categories were included: (iv) use of more than four bins, (v) use of a hat, and (vi) justification of the choice of the "better" class. For the fourth part of the protocol, the hat category also included the use of reference lines and numerical or percent values. Other categories recognised (vii) use of the mean and (viii) evidence of proportional reasoning.

Because the Data Cards protocol focused on the relationship among variables, the categories for analysis reflected the degree to which this occurred for the students using *TinkerPlots*. After noting (i) the list of variables with which the students engaged and (ii) the use of scaled plots or bins or a combination of the two, the categories recognised (iii) a focus on individual data values, (iv) the use of hats, (v) the relationships suggested among variables, and the (vi) identification of single variable characteristics (for example, "most").

It is clear that due to the interaction of the software with the tasks set for Sample C, generally more and somewhat different categories of response were possible for the *TinkerPlots* protocols than the earlier paper versions. In Samples A and B the focus was largely on the cognitive level of response rather than the interaction with the didactic objects, i.e., the graphs presented on paper or the graph-making materials available to the groups. For the *TinkerPlots* versions of the protocols, both the level of outcome and the interaction were noted. As noted earlier, however, it is not reasonable to compare levels across the software and non-software environments except insofar as the environments may support a particular level of response. Thus extracts and examples chosen from the earlier studies are included when relevant to the issues of interaction of students with the formats. A descriptive account is hence provided of similarities and differences of the usage made of the format for the two protocols.

5. RESULTS

5.1 Part (a) of Comparing Groups Protocol: Red and Blue Classes

Students in Sample A who saw this part of the protocol presented as in Figure 3, only had to recognise the scale on the graph, interpret the data as the number correct, and observe that the Red class had all scores higher than the Blue class. This was completed successfully by all students interviewed with the paper version of the protocol. Similarly all students in Sample C

using *TinkerPlots* reached the same conclusion. Because they had to interrogate the data, however, rather than observe a completed graph, several approaches were employed.

Due to the small data set (6 values for each class) three students in Sample C clicked through the cards or clicked on the icons in the plot to obtain the totals. One student did no further manipulation of the plot. Ten students created four bins, equivalent to the plot shown in Figure 7. Twelve students either used more bins (e.g., 6 x 2) or separated the data along an axis, with four including hats. Those using bins justified their decisions by stating the bin widths, 0-3 for Blue and 4-7 for Red, without knowing the exact scores. These students were not using visual clues in the same way as students who created a separated dot plot or read the pre-prepared graph; they demonstrated that grouping rather than knowing precise values was sufficient to reach a conclusion.

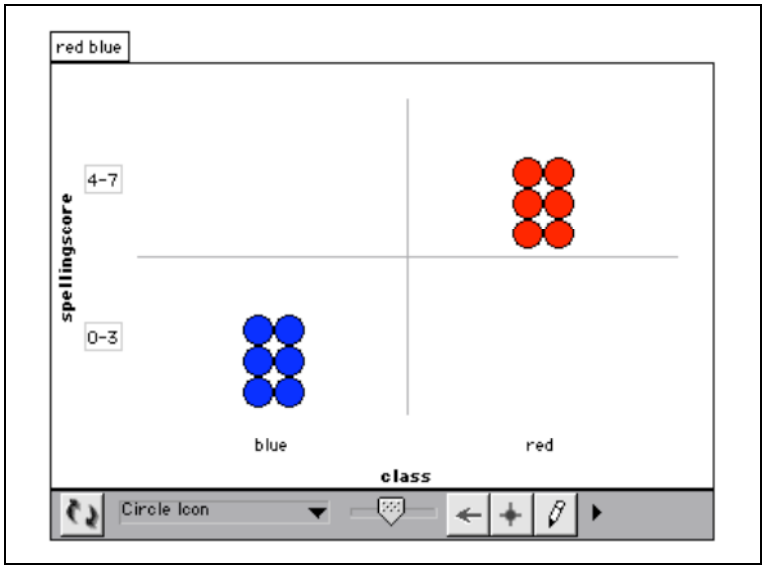


Figure 7: Use of 4 bins to compare the Red and Blue classes.

5.2 Part (b) of Comparing Groups Protocol: Green and Purple Classes

For the Part (b) of the Comparing Groups protocol, the ranges for the nine values in the data set were the same and it was a matter of observing the shapes of the graphs or adding the scores to obtain the totals, in order to decide which class had performed better. Both of these techniques were used by the students in Sample A with the paper version of the protocol, with younger students more likely to find totals and older students more likely to base their decision on the visual appearance of the graph.

Most students in Sample C using *TinkerPlots* were consistent in using the same approach to Part (b) as they had to Part (a). Similar reasoning was observed as for the paper protocol in comments like, "The Green have more in 6-8 and less in 3-5; Purple only one in 6-8." Some statements were incomplete, although reaching the appropriate conclusion. One student opened the Table and scrolled values, then plotted spelling score by student (number) and identified class by colour (see Figure 8), at first saying the classes were the same, then choosing Green because it had 3 sixes whereas Purple only had 1.

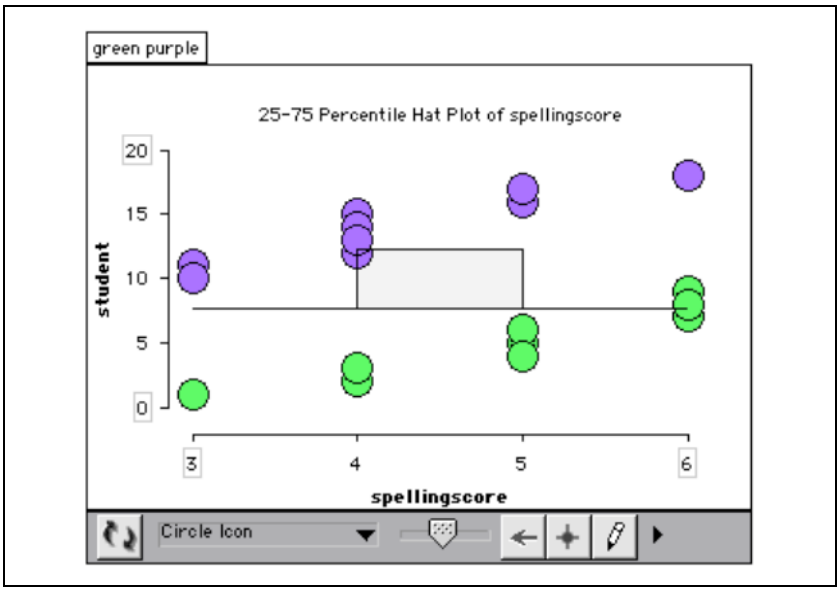


Figure 8. Plot displaying class (Green or Purple) by colour.

5.3 Part (c) of Comparing Groups Protocol: Yellow and Brown Classes

For Part C of the Comparing Groups protocol, the total scores for the two classes were the same, both distributions were symmetrical, but the Brown class had a wider range. Many students in Sample A when presented with the printed graph, immediately totalled the scores for each class mentally, pronounced them equal, and made no further comment. Occasionally means were calculated. Other students focussed on one of two distinguishing features of the two graphs, either the score of 7 for the Brown class, pronouncing it better, or the five 5s in the Yellow class, deciding it was better. Reasoning for choosing the Yellow class based on the 5s included the statement "more 5s," "more average," and "Yellow is more consistent." Some students, instead of calculating totals, studied the graph and used a balancing strategy to conclude that the classes were equal: for Yellow, 5 + 5 = 10, and for Brown, 3 + 7 = 10. Although finding the totals equal, some students went on to consider many features of the graphs, debating with themselves whether having a 7 was more important than more 5s, and discussing the range. Some compared the two graphs column by column and a few could not make a decision but were not prepared to declare the two classes equal. The following responses are from two grade 7 students.

S1: That one [Brown] goes up and down so it's a middling score, that one [Yellow] goes up and down but there aren't any 3s or 7s. So this class [Brown] has got the best result and also the lowest. This one [Yellow] got average results. [Counts both] I think they're both the same ... so equal.

S2: I think they probably scored about the same because people scored, in this one [Brown] they had a wider range of scores and there was two people scored 4 and 6 in both classes and that works out to be the same and the same with ... three people scored 5 in the Brown class and if three people scored 5 in the Yellow class then the other two 5s would add up to the same as the 3 and the 7.

In contrast to the students presented with the graph in Figure 3, the students in Sample C using *TinkerPlots* overwhelmingly (20 out of 24) chose the Brown class as better, whereas two chose Yellow and two said the classes performed the same. One student who said Yellow did better, again did not use any plots and looking at individual values decided that Yellow did not have a 3

and had “more altogether.” The other student said Yellow because there were more 5s. Of the two students who said the classes were equal, one used totals after separating the data into 5 x 2 bins; the other first used bins, then separated the data along the axis and made comments about Yellow being “taller” but Brown being “spread up the top but also down the bottom.”

Of the students using *TinkerPlots* who said Brown had done better, nine created four bins in their plots (see Figure 9) and made a decision based on Brown having more scores in its 6-8 bin. The other students created more bins, enough to separate the scores, or separated the data completely on two axes. These students made more descriptive comments, considering individual values, spread, the middle of Yellow, and the positioning of hats (4 students), but in the end chose Brown due to the presence of the score of 7.

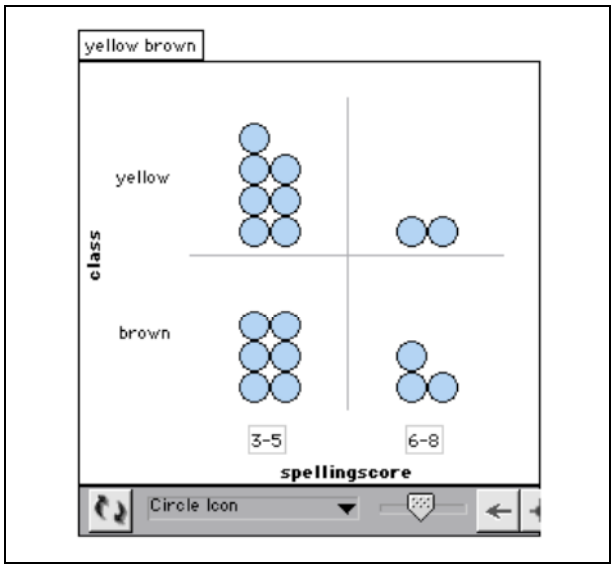


Figure 9: Use of 4 bins to compare the Yellow and Brown classes.

5.4 Part (d) of Comparing Groups Protocol: Pink and Black Classes

The final part of the Comparing Groups protocol considered two classes of different sizes to explore students’ understanding of how to address this issue. The question of interest here is not the percentage of students who chose the Black class but how students used either the graph presented in Figure 3 or the plot they created in *TinkerPlots* to make a decision.

For the students in Sample A presented with the printed graph, the visual appearance of the larger Pink class influenced many students to choose it. They focussed on the higher frequency of the 5s and 6s in the graph, not realising that the relative values of the scores decreased rather than increased the status of the Pink class relative to Black. Some students, while still deciding Pink did better, carefully compared the right-hand columns of the two graphs to see they were the same, then moved to the middle where the Pink columns were taller. Many did not appreciate, until told, that the size of the class could be a mitigating factor. Following the previous parts of the protocol, some students went straight to the total scores for the two classes, without considering the graphs.

Students in Sample A who chose the Black class as having performed better, showed various levels of engagement with the graph and how the representation related to their developing

proportional reasoning skills. Two responses from grade 6 and 7 students illustrate an intuitive appreciation of the class size and the meaning of the scores being higher or lower.

S3: The people in this class [Black] have done well for how many people there are, whereas this one [Pink]: more than probably about half of them are on the lower side, whereas this class [Black] hasn't got as many people on the lower side, more on the higher side.

S4: [Black] There are more children in the Pink than the Black … so you would have to count how many in each and then pick out what percentage of people got 6 and 7 … so more people out of Black, more percentage, got like the higher scores than in Pink.

Other students, usually a bit older, did not engage with the graphs at all, except to recognise a larger sample size.

S5: Well there's more people in this class [Pink] so you'd have to take an average for this one, you can't tell just by looking at the graph.

A few students considered global perspectives of the graphs as well as using their knowledge of means to create integrated responses.

S6: Okay, by averaging it Black scored better. They got 6.2 and the Pink class got 5.5. So even though they [Pink] had more people, they have more people who scored lower, like it kind of goes in an archish kind of shape [points to Pink], like they had more people score around the middle kind of range. Whereas bearing on the numbers in the class, they had more people score around the middle kind of area [Pink]. The Black class had more people score around the top kind of area. So averaging it Black still scored better.

In Sample C using *TinkerPlots*, no students used the means of the groups in making decisions about which class had done better. The grade 5/6 students had not been introduced to the mean in their classroom activities and although the grade 7 students had, only one put the means on the graphs and it was not used to make a decision. Only two students restricted their plots to four bins (see Figure 10); both based a decision on there being more Pink in the 5-9 bin.

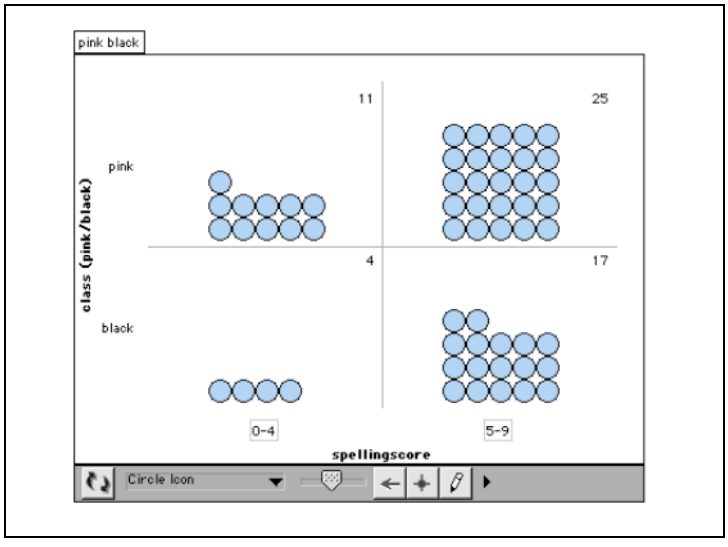


Figure 10: Use of 4 bins to compare the Pink and Black classes.

One grade 7 student produced the plot on the left in Figure 11 and displayed the following intuitive reasoning.

S7: It kinda looks like this one [Black] did better 'cos most of them are up there [5-9] and there is only 4 up there [0-4] … Pink could be better because it's got 25 up there and 17 there [comparing 5-9 for Pink and Black] so there is more that has 5, 6, 7, 8, 9 than Black. But Black got less down here [0-4] because it's got less people than this [Pink] class … You never know if you have the same [number] in both classes then you might've got a more even thing … Black did better actually because this one's [Black] got less people in it but 11 got it wrong, got like 1, 2, 3, 4 and only 4 got it here [0-4] and 17 got it there [5-9] but 25 got it there [5-9] only because there's more in the Pink class.

When prompted by the interviewer she then separated the data and was asked if this changed her mind about the Black class doing better.

S7: This one [Pink] looks like it's got more high scores than this one [Black] … I don't think I would change my answer because I would just keep it the same because I reckon Black did better because it has less kids but it has more in the higher class [score] than the lower score.

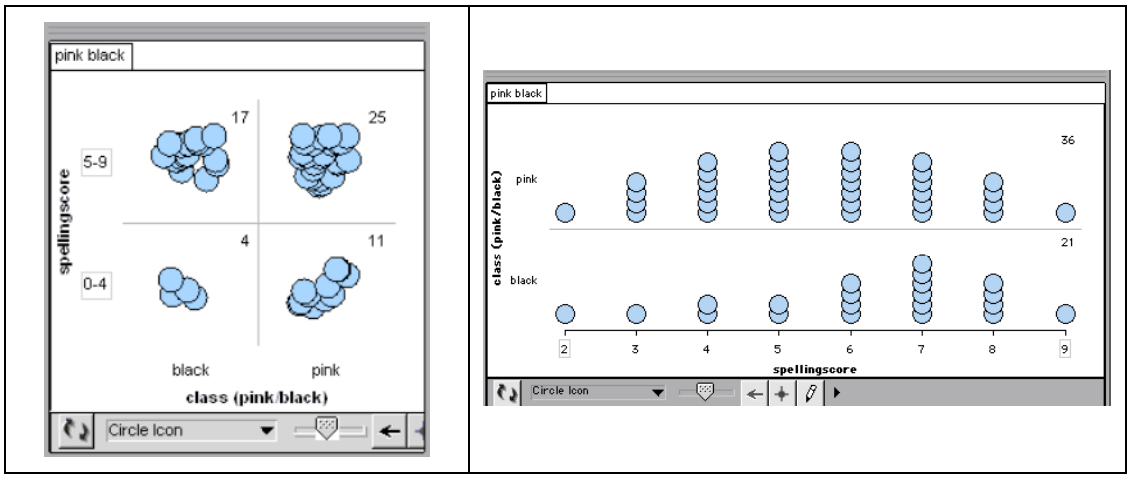


Figure 11: Plots associated with intuitive proportional reasoning (S7).

Another grade 7 student created two bins by class to see that the sizes of the classes were different, then created a colour-coded stacked dot plot for the two classes combined (see Figure 12). He then chose the Black class because "Black has more on these [higher] numbers and Pink has more down on the smaller numbers."

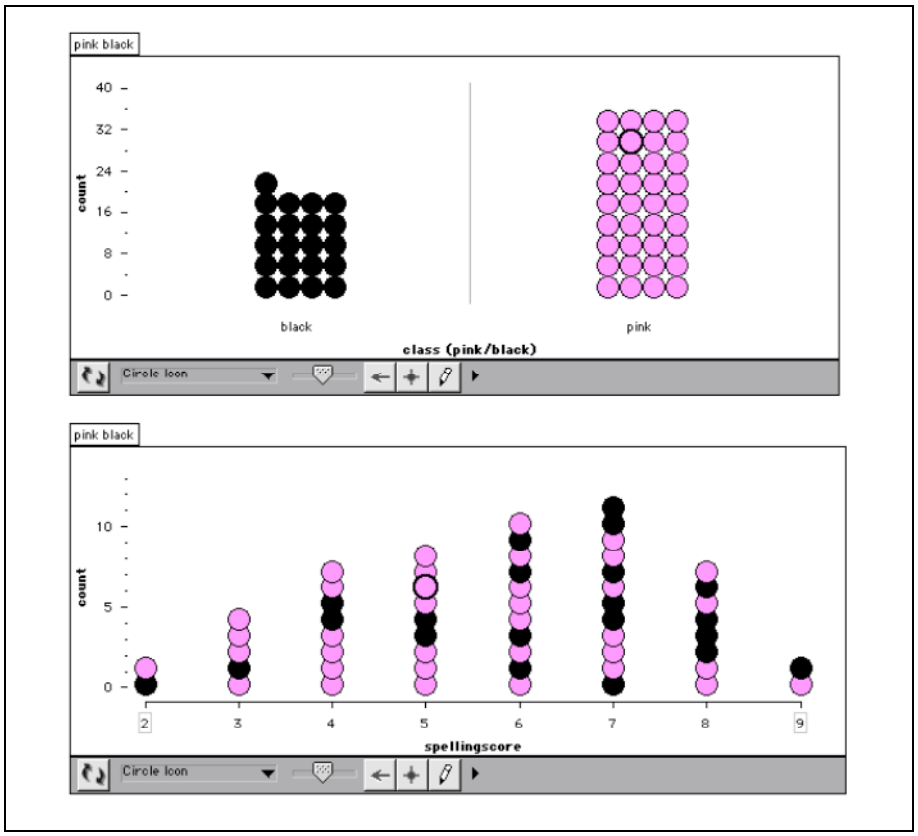


Figure 12: Use of bins to count group sizes and of class colour in the stacked dot plot.

The rest of the students in Sample C either had individual bins for all spelling score values for each class or separated the scores along axes as stacked dot plots. Half of the students introduced hat plots to their stacked dot plots (see Figure 13); some were used effectively but others were not. At times the hats appeared to help students distinguish across the range of scores. One student, for example, commented, "Black is the least down here [low scores] but Pink has lots. Pink is half and half. Black is 75% up here [high scores] and 25% down here [low scores]." Several students, even when alerted to the sizes of the classes, did not realize that this was a significant issue. Typical comments along this line included "Pink has more numbers and students – they went better." A few students had not absorbed or at least not retained a complete appreciation of the information conveyed in the hat plot in relation to the decision required: "Pink better because they have more in the 50% part of the hat; the Pink hat is bigger." Overall once students separated the data along the axis or in bins for single values of spelling score, their success rate was little different than for the students in Sample A interviewed with the paper version in Figure 3.

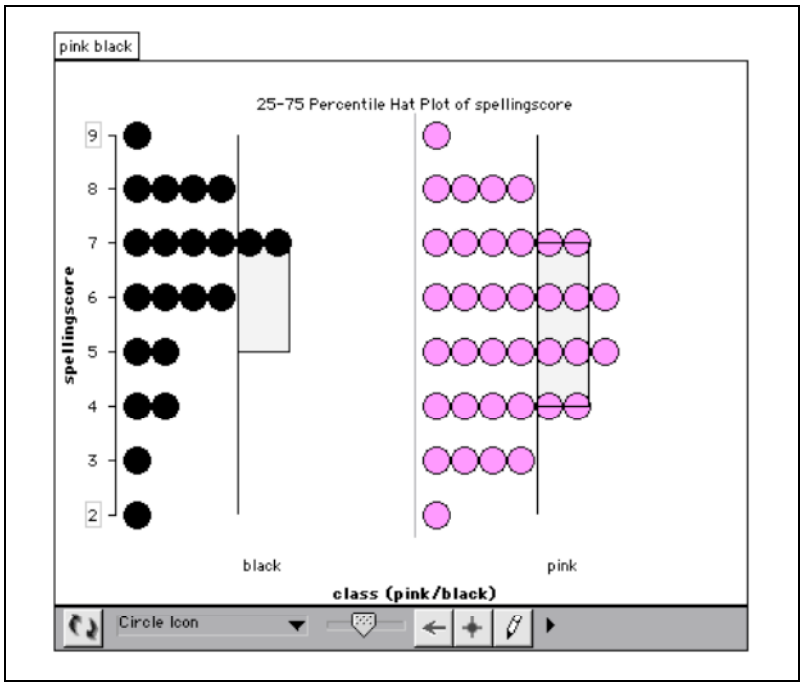


Figure 13: Use of scaled plots with hats.

Table 1 summarises the differences observed over the four parts of the Comparing Groups protocol in relation to the affordances offered in the software and non-software settings for students to display their understanding of the procedures for comparing two groups, for students to display different understandings, and for time to be conserved. In each setting the data were identical, as were the questions asked of the students. The opportunity to display cognisance of the importance of unequal groups was considered to be similar in the two contexts, as was the difficulty of appreciating proportional reasoning. The potential of the Comparing Groups protocol to expose student levels of understanding of ways of drawing comparisons of two groups of equal or unequal size hence appeared just as strong with *TinkerPlots* as without.

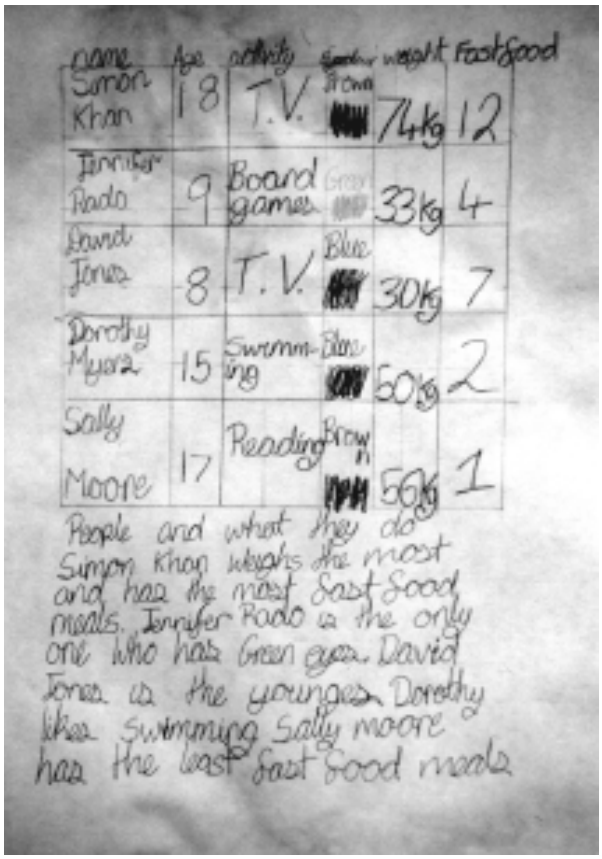
TinkerPlots (Sample C)	Paper (Sample A)
Students had to process raw data.	Summary of data provided in graphs.
Variety of representations possible. (i) Possibility to stop separating when enough information displayed. (ii) Possibility to use one graph with colour-coding for groups.	Single graphical representation to be interpreted.
	Students more likely to compute totals for groups.
Data accessible in data cards OR in a table.	Data only available in graphical format.
Use of bins sometimes hid individual values.	All data values visible in graphs.
Modest use of hats to explore middle.	Modest use of means to explore middles.
Time involved sometimes longer.	Time involved sometimes shorter.

Table 1: Summary of Differences in Affordances in *TinkerPlots* and Paper Formats for the Comparing Groups Protocol.

5.5 Data Cards Protocol

The Data Cards protocol was much more open-ended than the Comparing Groups protocol. Students were not asked specific questions but were asked to form hypotheses and to create graphs of evidence to support their claims. Students in both Sample B and C spent time flicking through the cards to find interesting cases. In fact it was an interactive process with graphs being created at times to assist in forming hypotheses. Due to the ease with which students could create plots in *TinkerPlots* compared to by-hand with paper-based materials, more graphs were created with the *TinkerPlots* by Sample C students than were produced on posters by the classroom collaborative groups in Sample B.

For the students working in classroom groups in Sample B, discussion fairly quickly and naturally considered the boys and girls in the data set (there were 8 of each) and a gender variable was defined and used in relation to the other variables. Of interest was how the representations created by the students reflected the associations present in the data set allowing them to display their level of understanding of the task. There was a total of 27 representations presented by the classroom groups, varying from one for a group to six. Students were only given the data on physical cards and the group with only one representation created a single table with all of the data collected together. Three other tables were presented, along with other graphs; these displayed information of interest to the students. One of these is shown in Figure 14. Bar graphs showing frequencies for single variables were the most frequently used graphs, as shown for example in Figure 15. There was also a single pie chart for eye colour. Eight of the graphs showed relationships among two or more variables. The graph in Figure 16 shows a graph claimed to be a bar graph that actually is an idiosyncratic scattergram, whereas the representation in Figure 17 relates three variables together. Although students had three sessions to explore the data, form hypotheses and create representations, their skills at producing graphs and their backgrounds meant that the number of types of graphs created was limited.



name	Age	activity	Eye colour	weight	Fast food
Simon Khan	18	T.V.	Brown	74kg	12
Jennifer Rado	9	Board games	Green	33kg	4
David Jones	8	T.V.	Blue	30kg	7
Dorothy Myers	15	Swimming	Blue	50kg	2
Sally Moore	17	Reading	Brown	56kg	1

People and what they do
Simon Khan weighs the most and has the most fast food meals. Jennifer Rado is the only one who has Green eyes. David Jones is the youngest. Dorothy likes swimming Sally moore has the least fast food meals

Figure 14: Table displaying data of interest to the students.

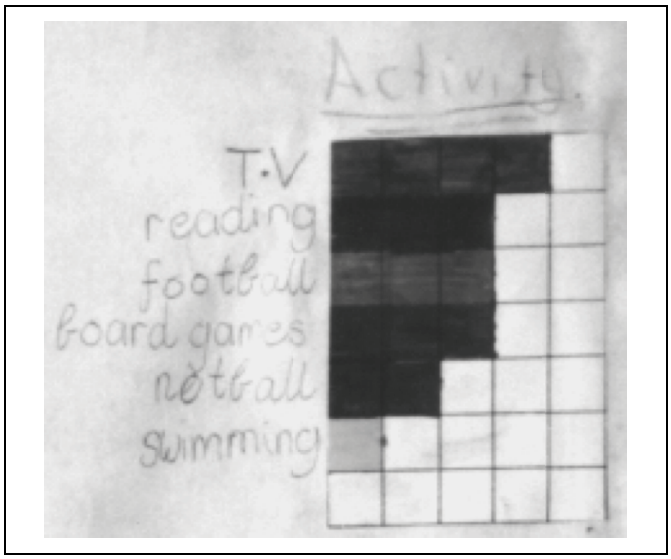


Figure 15: Bar graph created for one variable.

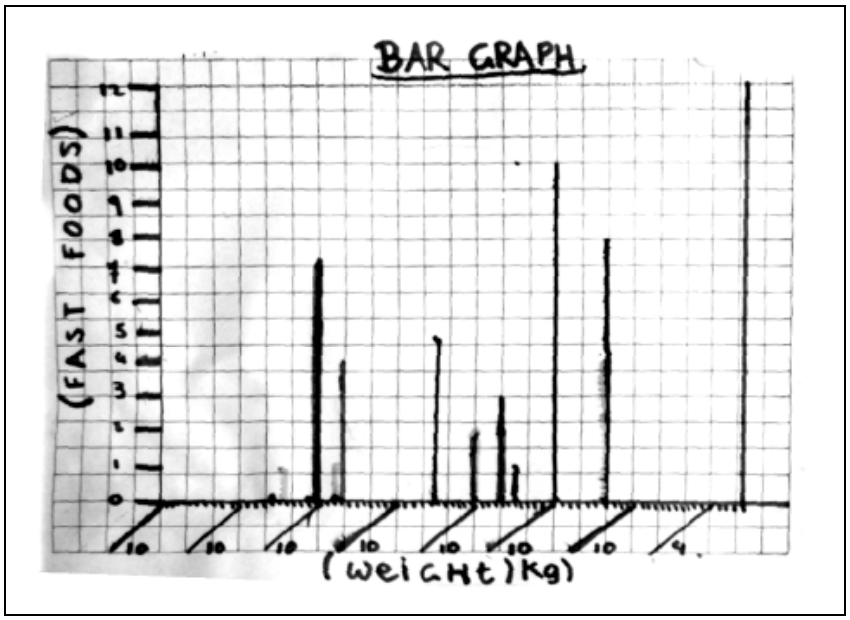


Figure 16: Idiosyncratic scattergraph created for two variables.

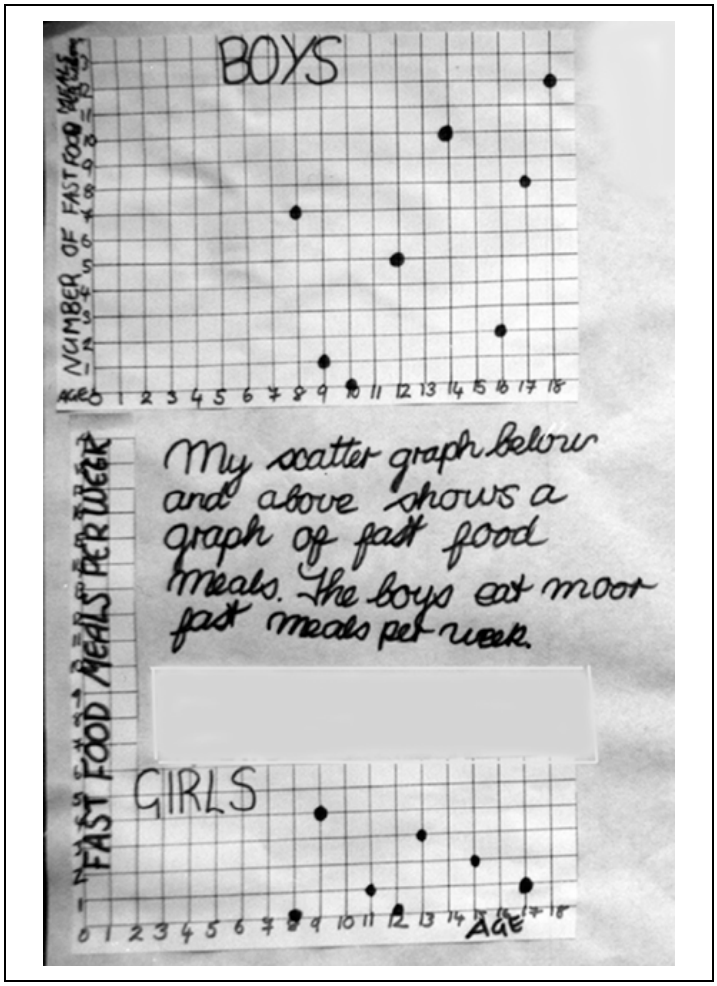


Figure 17: Three variables in a poster.

For Sample C students in the *TinkerPlots* interviews, finding the oldest and youngest individuals and who had the most fast foods and was the heaviest was of interest to them. Some students looked at many cards. With data in bins, many students checked out values by clicking on the icons on the plots (to display the associated data card) and discussed values for other attributes in relation to the attributes on the axes of the plots. This information was used to make hypotheses for a few cases with the cell; for example, one grade 7 student went through the bins on a favourite activity x age plot (with two cells for age) making comments across the ages and activities but had difficulty suggesting any trends for more than a few individuals at a time. Another student (grade 5/6) when asked to form a hypothesis about weight and favourite activity, described individuals across the activities: "Wow, this person here, weighs 75. They must have eaten a lot of junk food. These people around 30 and 35, play netball … He or she watches TV and is still about 30 kg. … This person swims." When asked for an hypothesis about fast foods per week, the student gave advice instead: "I reckon they should cut down. And eat some vitamins and stuff."

For variables that could be separated completely along an axis there was a tendency for students using *TinkerPlots* to leave them in bins. Sometimes this was not a problem because the relationships were clearly shown but at other times the small number of bins used meant that it was difficult to hypothesise about associations. Not very many students went as far as creating a scattergraph with two scaled axes.

Probably because of the time constraints and the lack of fellow students with whom to discuss the data cards, no students using *TinkerPlots* naturally suggested gender as a variable. For the grade 7 students, the interviewer often suggested the definition of a gender variable toward the end of the interview. Some students took up the suggestion and created a variable, sex, and made some hypotheses based on it.

Although asked to consider relationships among the variables, some students could not move past describing single variables. The statements made included the following.

- The most popular eye colour is blue.
- Board games are the lowest favourite activity and netball is the middle.
- Most people weigh 50 [60] to 30 kg.
- There are more in the TV group [of favourite activity].
- More people in the 0-39 [kg] group.

For students who defined the gender variable, it provided interesting plots with bins that were the equivalent of 2-way tables. The plot in Figure 18 shows the relationship of favourite activity and sex in a 6 x 2 bins arrangement. Two plots in Figure 19 illustrate the way bins were often used with a scaled variable. Although no students continued to separate the fast food variable, these plots were sufficient to hypothesise about the relationship of gender and fast food meals per week.

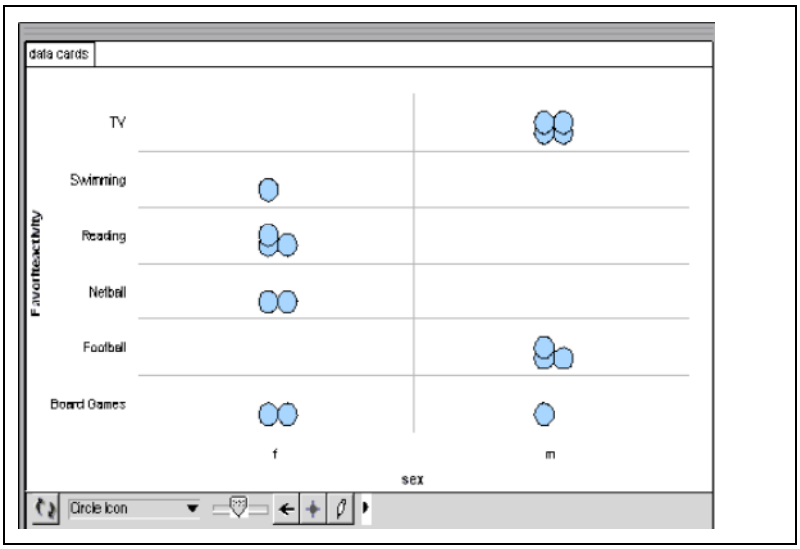


Figure 18: Use of bins to consider favourite activity and gender.

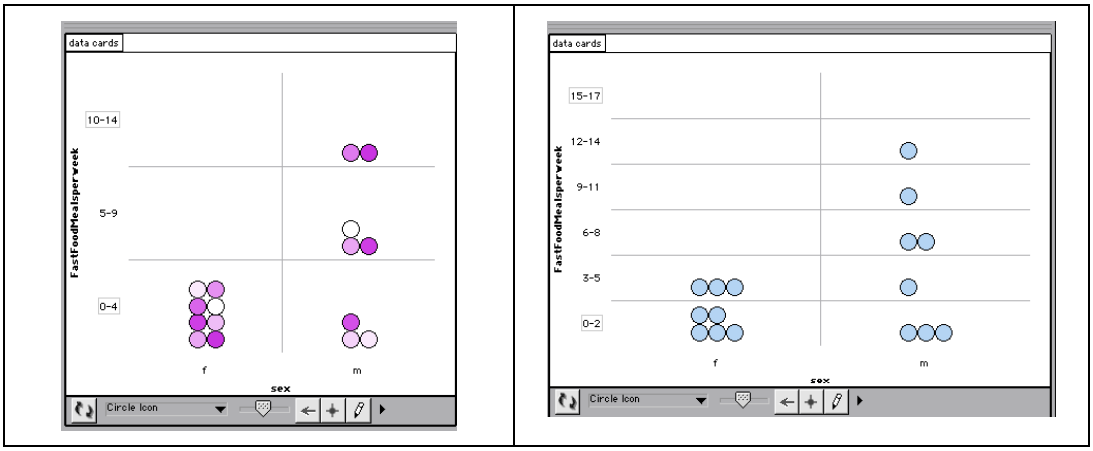


Figure 19: Use of bins to consider fast food meals per week and gender.

Considering fast food consumption and favourite activity produced several variations on using bins and a scaled axis. Figure 20 shows three of these. On the left the separation into two bins for fast foods is enough to make a hypothesis about fast food consumption and watching television. The middle plot shows the maximum number of bins and hence this representation holds the same information as the plot on the right, which is fully separated and includes hats.

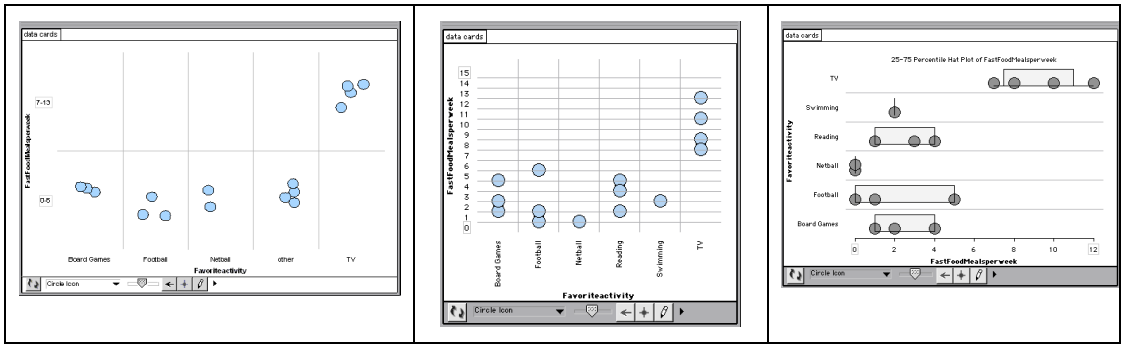


Figure 20: Use of both bins and a scaled axis to consider fast food meals and favourite activity.

The use of various plot representations for two measurement variables is shown in a group of four plots created by different students to explore the relationship of age and weight. The top plots in Figure 21 show 2 x 2 bins and 5 x 2 bins, the latter with the number in each cell labelled. The bottom left plot separates the weight variable and the bottom right plot separates both to create a scattergraph. All of these representations are effective ways of showing the association of age and weight because it is quite strong.

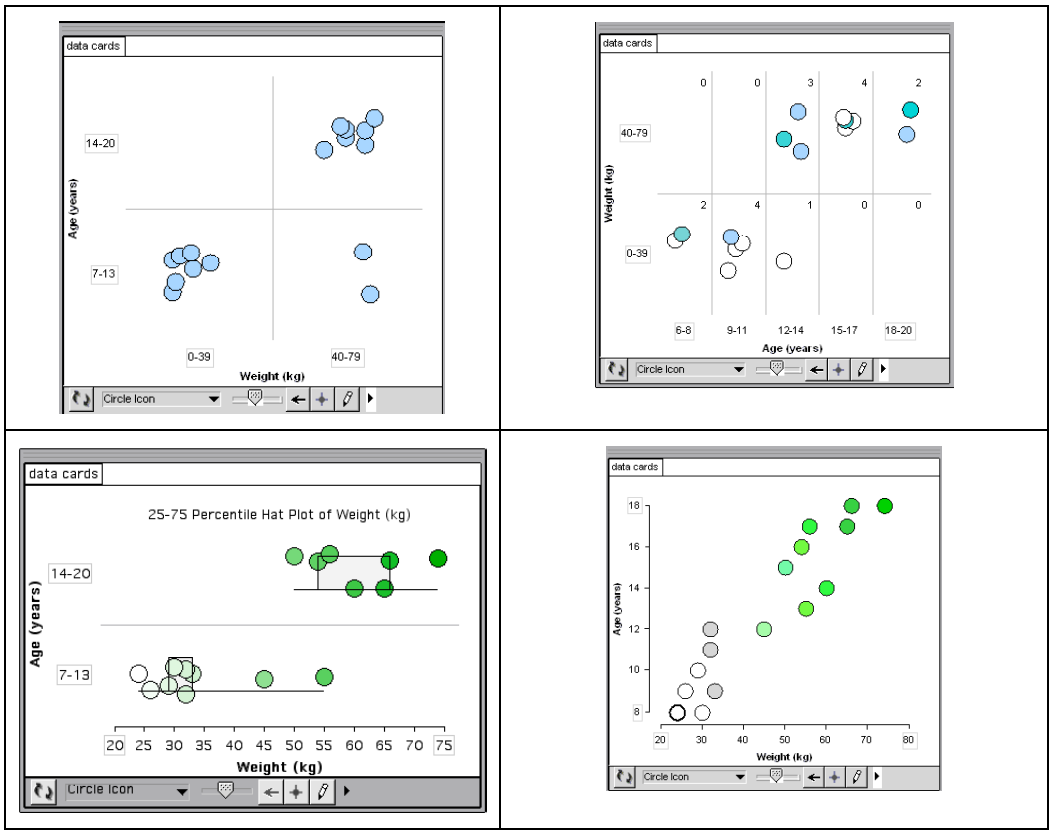


Figure 21. Different representations for age and weight.

A similar range of representations was presented in considering the association of fast food and age and fast food and weight. Figure 22 shows the progression for fast food and age, where no one created a scattergraph with two scaled axes. The plot in the middle of Figure 22 with 6 x 5 bins to separate the data partially is quite close to a scattergraph in showing the association between the two variables. The separated data for fast food meals per week in the plot on the right is floating (not stacked), which means that the details of the arrangement of the icons could be misleading if considered other than in relation to their position to the left or right across the plot. Four plots for fast food and weight in Figure 23 illustrate the value of the scattergraph for this association. It was more difficult for students to suggest a relationship between weight and fast food consumption for the data presented in bins. The variety displayed in Figures 22 and 23 also includes choice of axis for the display of attributes.

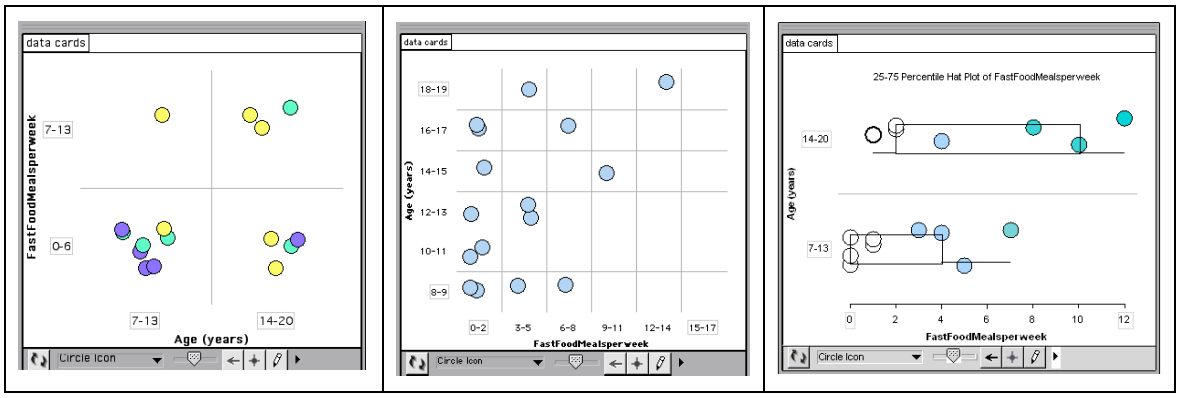


Figure 22: Representations for fast food meals per week and age.

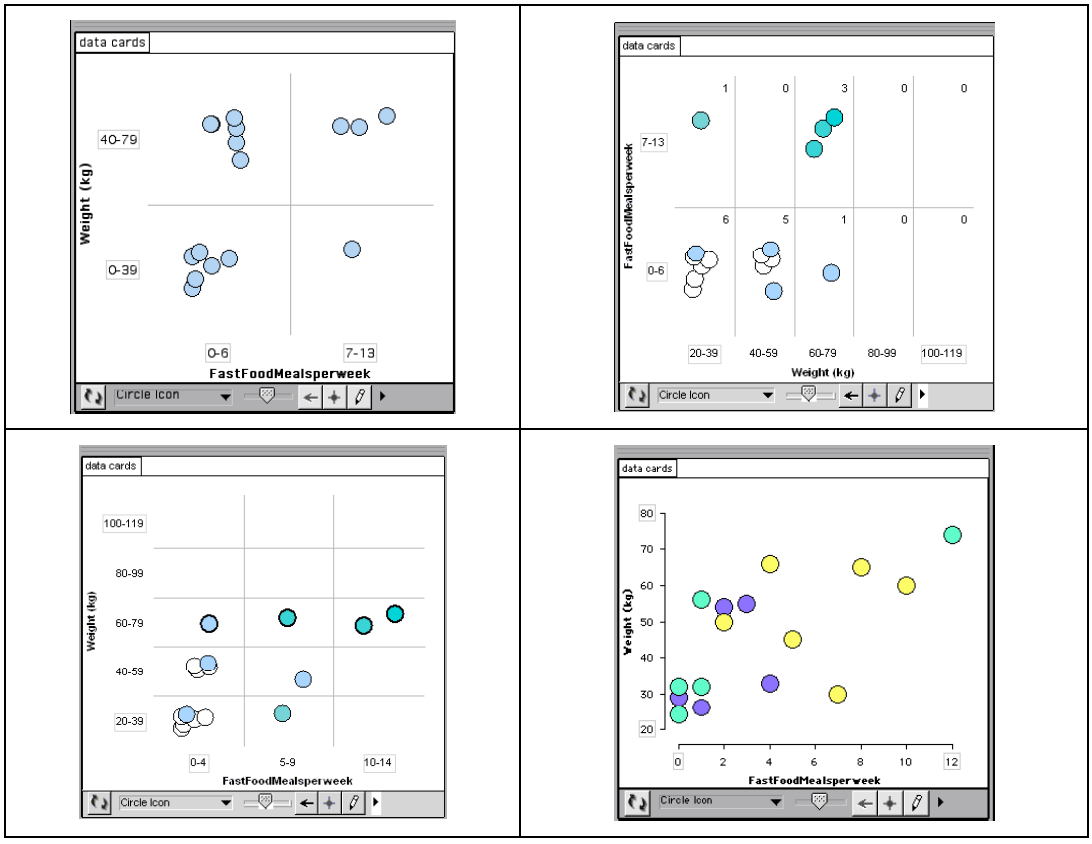
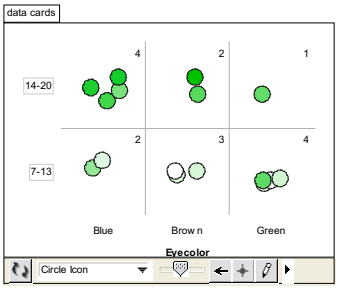


Figure 23: Representations for fast food meals per week and weight.

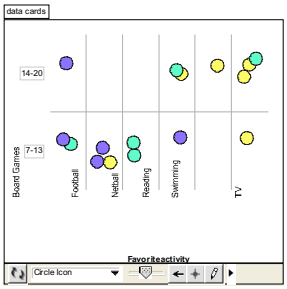
In considering potential hypotheses for the relationship of variables in the data set, most students in Sample C created plots before suggesting hypotheses. One grade 7 student, however suggested the following hypotheses after looking at the variables as listed on the individual data cards: "maybe older people eat more fast foods than younger people; people that weighed more are more older; males weigh more than females." He then looked at many cards and described them before creating seven different plots to confirm these hypotheses and also look at favourite activity. A grade 5/6 girl wanted to look at age, weight, and fast food together but could not create a plot for all three, so used the cards by clicking on the icons in a fast food x weight plot to

claim that people who do more exercise have less fast food (fewer meals) and less weight. Another grade 5/6 student looked at individual cases by clicking on icons on the weight x age graph and concluded that less fast food was associated with less weight.

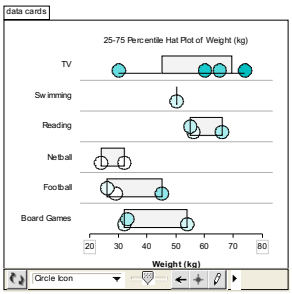
Many students gave more individual explanations than general trends and some of the claims were dubious based on the data available. Some of these are shown in Figure 24, where student quotes are presented under the plots produced. Some of the comments were associated with clicking on icons as well as considering the attributes on the plots. The comments by S11 were made by colouring the icons in the plot with the eye colour attribute.



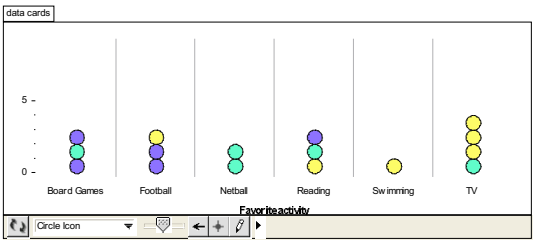
S8: Well, if you're in this group there is more people from 14-20 that have blue eyes than all the others.



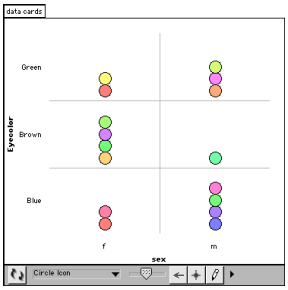
S9: The younger people are more fitter and they get more active than the 14-20 year old kids.



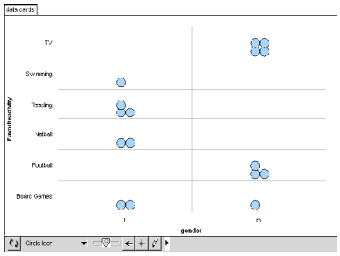
S10: The people that play more games and sports that you move around and stuff they're more healthy, the people that were playing, like, their favourite activity was TV, they're not as healthy and they are eating more fast food. [What do you mean by more healthy?] They're eating less and they have got less weight on them plus they are more active.



S11: [icons were colour-keyed to eye colour] Green eyed people like netball, actually they are brown eyed. Blue eyed people like TV better, green eyed people like football better, and board games.



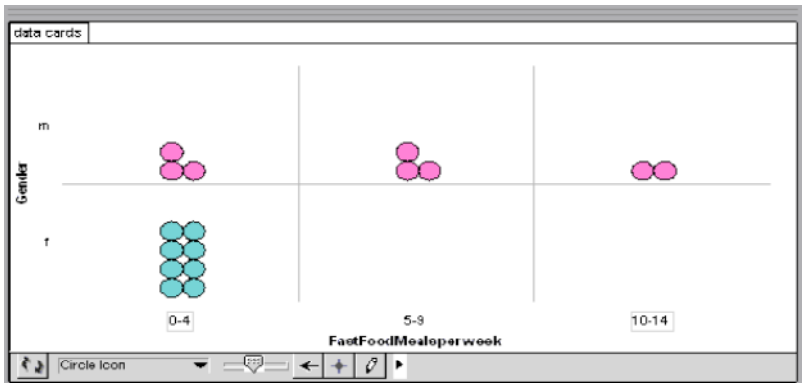
S12: Girls have brown eyes, boys have got more blue.



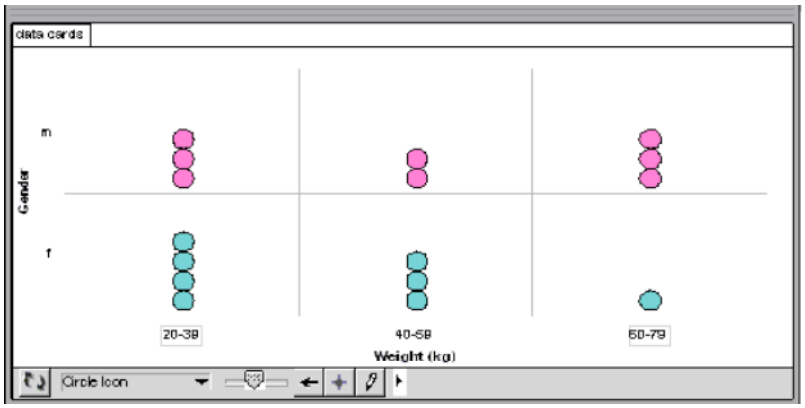
S13: The boys like to do some lazy things and football.

Figure 24: Students' implications from their graphs.

Students generally did not make very many comments specifically associated with variation when discussing their hypotheses. One grade 7 student, who defined a gender variable and produced the plots shown in Figure 25 drew the conclusions presented below each plot. The last hypothesis from the data did not fit her expectation, because she liked football, and she stated the need to interview more students to obtain more data.



Males eat more fast foods than females. Males are more spread and females are in one group.



Males are heavier than females. *Boys eat more fast foods so that makes them more bigger.*

Boys and girls like completely different things.

Figure 25: Representations for gender and three other variables.

Table 2 summarises the similarities and differences in affordances for the software and non-software settings of the Data Cards protocol. As the protocol was more open-ended than the Comparing Groups protocol, more observations were made of similar and distinguishing features. Overall the protocol itself appears well suited to exploring student understanding with efficiency benefits if time is a factor in assessing learning outcomes.

Similarities	Differences	
	TinkerPlots (Sample C)	Paper (Sample B)
Intuitive initial hypotheses.	Gender variable not naturally added to cards.	Gender variable came naturally from physical cards.
Bins functioned rather like tables.	Frequent use of bins in plots.	Conventional table and graphical forms.
Some students only considered one variable at a time.	More graphs produced.	Fewer graphs produced.
Variety of representations produced; flexible choice of axes.	Quicker to produce different plots.	Time-consuming to create graphs by hand.
Some students continued to focus on individuals.	Easy to identify individual cases by clicking on icons.	Need to find physical cards to identify individuals.
Same associations considered, but few produced scattergraphs.	Overall faster to complete.	Overall slower to complete.
Lack of sample-population links.		

Table 2: Summary of Similarities and Differences in Affordances in TinkerPlots and Paper Formats for the Data Cards Protocol.

6. DISCUSSION

The affordances of *TinkerPlots* for exploring student understanding of statistical concepts fall into three categories: flexibility of representation, speed of analysis, and exposure of levels of understanding. These are discussed in turn, followed by the issues that arose in relation to researchers' need to be cautious, the wider use of *TinkerPlots*, and the potential for further research.

6.1 Flexibility of Representation

The major affordance that *TinkerPlots* offers over print versions of tasks such as the Comparing Groups protocol, is the flexibility that it gives students to create their own representations of data sets to display their understanding. Although certainly aware of graphs similar to those in the print version of the protocol, many students produced different plots. For the initial two data sets, these were often basic, for example using four bins, and were sufficient to answer the questions

appropriately. The procedure that is available in *TinkerPlots* for initially displaying data along an axis, beginning with two bins and then requiring an icon to be dragged to the right to separate the data further, means that students may stop dragging at a point where the data are able to tell the story they want to relate rather than continuing to separate the data completely. This is illustrated in Figure 26 for the complete data set for the Pink and Black classes. Also illustrated in Figure 26 is the affordance offered by *TinkerPlots* to distinguish values of an attribute by colour; this was used on several occasions by the students interviewed.

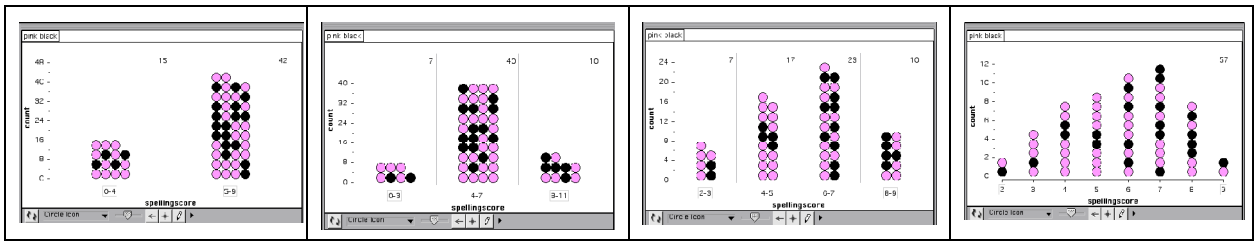


Figure 26: The steps in separating data along an axis with *TinkerPlots*.

Another aspect of the flexibility for students in their presentation of plots was the tendency of many to drag the variable of interest, for example spelling score or fast food meals, to the vertical, rather than the horizontal axis. Given that much of the instruction students would have had about drawing graphs would have been based on starting with a horizontal axis, this may be surprising. Whether it could be related to dragging the variable to the "nearest" side of the plot is unknown. This could be a question asked in further studies.

As well as the flexibility of graphical representation, *TinkerPlots* also offered both the set of data cards and the table format containing the data. More of the students in this study looked at the data cards, either by clicking through them by number or by clicking on individual icons, which brought the particular card to the top of the virtual pile. Some, however, did scan through the table, especially for the Data Cards protocol. Having the data available in different formats meant that student styles of expressing their understanding were catered for within the package itself. In the paper versions of the protocols students had to conform to the conventional representation presented to them.

6.2 Speed of Analysis

For the Comparing Groups protocol, the time taken to answer the questions by Sample C students using the *TinkerPlots* format was marginally longer than for students in Sample A because the students had to create their own plot or examine the data. The affordance of observing students' preferences for their method of analysis, however, outweighed the time issue. For the Data Cards protocol, however, the amount of time saved was significant in being able to plot attributes quickly, confirming, disregarding, or creating hypotheses about relationships among the attributes. Most of the Sample C students discussed the same associations of attributes as were considered by the Sample B students working in groups and by the few students interviewed individually with the physical data cards. In terms of discovering the students' appreciation of possible relationships and ways to portray them, the interaction of a group setting did not appear necessary. The time that was spent by students working in groups to create attractive graphs with concrete materials was skipped for the interviewed students. For the explicit purpose of exploring students' understanding of finding evidence for associations of variables, the *TinkerPlots* format was much more efficient.

Part of the issue of time in using *TinkerPlots* to create plots, for example for the Data Cards protocol, is also seen in comparison to the time wasted by the students in Sample B working in groups when mistakes were made in creating graphs, with students throwing away graphs and starting over. Some students working in groups also prepared drafts before creating final versions for display. In contrast, a decision could be made easily to delete a *TinkerPlots* plot and replace it with another. Although learning styles may influence student preference for these two types of analysis and presentation, given the trend for using sophisticated technology to summarise data across society, it would suggest that exposure to a learning experience appropriate for the middle school is a good investment of curriculum time. The researchers would appear to benefit in terms of flexibility for students to display their understanding.

6.3 Exposure of Levels of Understanding

In using any research methodology to explore student understanding, the aim is not only to discover unusual or brilliant approaches to solving a problem, but also to document incomplete understandings or misconceptions. These often occur in the form of over generalisation from simpler cases to more complex ones. Observing such examples during student interviews then informs teachers, who can look for similar instances and address them as soon as they occur in the classroom. An example of this for the paper version of the Comparing Groups protocol used in Sample A was the tendency for students who had used totals (of scores for each class) to compare the first three pairs of classes, which were of equal size, to continue to use totals for the Pink and Black classes, where the sizes of the classes were different. This tendency did not occur for the students in Sample C using *TinkerPlots*, mainly because of the visual cues obtained from using bins as a way of representing the data. The overuse of bins, however, without distinguishing individual scores, for example for the Yellow and Brown classes, led to making a decision without considering mitigating values that made the total scores equal. The affordance shown in Figure 26 hence needs to be tempered with caution about being sure students are aware of considering all possible representations. As well, the issue of the number of scores in bins, without considering the proportion of the class, confused some students using *TinkerPlots*, in a similar way visually, to the students using the paper format.

It is interesting to note that the use of bins by Sample C students when they could have separated the data along an axis completely did not seem to disadvantage students where answering the Data Cards protocol compared to students in Sample B. The fact that Sample B students did not use an equivalent of bins is likely due to all data appearing in integer form and students having experience in plotting individual values along labelled axes. To have grouped data into categories would have seemed a more complex unnecessary operation based on their previous learning. For the Sample C students, having been presented with bins initially by *TinkerPlots*, meant that in some cases dragging the icons to separate the data was the more complex unnecessary operation. For these students, they could already *see* the relationship displayed (see examples in Figures 19 to 23).

Not all students in Sample C used hat plots in their interviews with *TinkerPlots*, although hats had been introduced and used previously. As noted it has been claimed that hat plots are easier for students to interpret because they simplify the information held, focusing only on the middle 50% rather than splitting the middle 50% with the median, where a smaller resulting quarter represents more tightly packed values rather than a smaller number of values. There were, however, a couple of students who used hat plots to compare two data sets and commented that a wider crown of the hat meant it necessarily had more values in it than a smaller hat crown. These

students had hence not appreciated fully the importance of the part-whole proportional relationship displayed by the hat and the implications for the spread of the data rather than the actual number of values. Actual conversations with the students during the interviews displayed these difficulties whereas they might have been lost in the classroom. As well a few students who used hat plots for the Pink and Black classes had difficulty interpreting the significance of the overlap of the crowns of the hats.

Because it was an interview setting using *TinkerPlots* and the students in Sample C were experiencing the data and variables in the Data Cards protocol for the first time, the interest in individuals and their characteristics is not surprising. The students in the classroom setting were observed to have a similar interest in the data in individual cards during their first working session. For some students in both settings this continued across the exposure to the protocol because they had difficulty taking in more than a few individuals (in a bin for example) or the relationships among variables on a larger scale. This interest is typical of students whose developmental level is characterised by considering single elements associated with a task rather than being able to process multiple elements and relate them together. Students operating at this level were observed just as well in the *TinkerPlots* setting as in the classroom setting.

Another issue related to level of performance was the tendency of some students in Sample C in the interviews with *TinkerPlots* to digress to giving advice about how the individuals on the data cards ought to behave, for example "get more exercise" or "eat more healthy food." This is typical of students finding it difficult to engage with a task at even the single element level. Similar comments were observed during the discussions of the students in Sample B working in groups but they did not appear on the final posters produced.

6.4 Cautions

Less experienced users of *TinkerPlots* may not consider defining other variables, such as gender, which was possible from the names provided for the Data Cards protocol. This was the case for the *TinkerPlots* students in this study, who had previously entered attributes, including gender, from their class data based on names but had not defined further attributes from a data set already presented to them within the software. Asking a prompting question about gender may result in students realising that they could define another attribute.

There were times during the interviews when the interviewer knew that the student had been exposed to and used a certain *TinkerPlots* tool that was appropriate at that point but the student had apparently forgotten it. Occasionally the student was reminded, said "aha" and used the tool then and later. At other times a hint to use a tool brought forth recognition but not the facility in appreciating what the tool offered. Hence in using *TinkerPlots* as a research tool to explore student understanding, it is recommended that a check list be used at the beginning of the interview to assess which of the *TinkerPlots* tools are familiar to the student and which are not. It is unlikely that teaching unfamiliar tools at this point would be feasible but feedback for the teacher in terms of further classroom experiences could be useful.

For both settings, with and without *TinkerPlots*, for the Data Cards protocol, the students had been exposed to the concepts of sample and population. For the classroom group-work setting (Sample B), the teacher-researcher had asked questions about the population of Australians and of people their age and for example their favourite activities and eating of fast foods per week. The cards were then introduced as a sample from a school age population. For the students in Sample C using *TinkerPlots*, classroom activities for other data sets had stressed the relationship

of samples to populations. In both settings, however, these concepts did not transfer in a strong fashion, either to the posters produced or to the discussion related in the *TinkerPlots* interviews. Although this issue is more closely related to classroom teaching and experiences than to the use of *TinkerPlots*, it is one of the observations that arose in the *TinkerPlots* interview context. The software itself would not seem to provide a solution.

6.5 Wider Use of *TinkerPlots*

This report has focused on using *TinkerPlots* as a research tool to explore student understanding in relation to comparing data sets and exploring a data set for relationships among variables. Having claimed that *TinkerPlots* saved time in relation to the Data Cards protocol, it must be acknowledged that time had been previously devoted to students becoming familiar with the software and exploring some other data sets. The overall investment of classroom time must then be gauged by the teacher in relation to the long term benefits of introducing the software or not. The fact that beyond assessing student understanding, using *TinkerPlots* also means that plots can be embedded in reports prepared in a similar fashion to the posters prepared by the collaborative groups, is likely to be an important factor in deciding to use *TinkerPlots* in the classroom.

Throughout this report there have been occasional comments in relation to classroom experiences, related both to experiences with the software, for example with the use of bins and hats, and to wider conceptual issues, for example related to samples and populations. It is clear that in exploring students' levels of development of statistical concepts, assessment is the outcome. How the assessment is used beyond the researchers' realm is a more open question. It is hoped that feedback provided to teachers would help in planning classroom interventions that would assist students to progress in three areas: facility with the tools in *TinkerPlots*, appreciation of the sample-population relationship behind many statistical questions, and personal development to higher levels of performance on similar tasks.

The great variety of plots produced by the students interviewed with *TinkerPlots* points to opportunities for classroom discussion if students are allowed to create their own plots and then share them with each other. This was not possible in the interview setting because of the objectives of the research. It would be a shame, however, to miss this opportunity in the classroom, for example by a teacher "instructing" the class on "how to produce a plot." Interpreting each others' plots would be excellent experience for both the reader and the creator of plots.

More specifically, it would be of interest to propose in a classroom context, a mixture of the settings for the Data Cards protocol. Students working in groups of two or three with *TinkerPlots* could be asked to spend several sessions producing a report using *TinkerPlots* supplementing their plots with text boxes that included their observations, suggestions and hypotheses based on the data set supplied, or even including more research of their own.

6.6 Further Research

The two interview protocols adapted for use with *TinkerPlots* had been shown to be valid in displaying levels of student development of understanding in earlier studies (Chick & Watson, 2001; Watson & Moritz, 1999). Being based on actual numerical data sets they were appropriate for translation into *TinkerPlots*. Not all concepts in statistics are so readily translatable. New versions of *TinkerPlots*, however, including a probability sampler (Konold et al., 2007; Konold

& Lehrer, 2008) offer new possibilities for exploring students' probability understanding via in-depth interviews (e.g., Ireland & Watson, 2008).

The outcomes of this study suggest that there is potential in further development of settings and prompts that would explore students' in-depth understanding of measures of average and of hat plots, with these leading to appreciation of box plots. The issue of box plots being displayed without the accompanying data values has been noted as a difficulty for student understanding (Pfannkuch, 2006) and the *TinkerPlots* facility to "hide" and "show" data values easily may provide a method of exploring student understanding in this area. As noted there is also the possibility of working closely with classroom activities, such as happened in the work of Watson and Chick (2001) with the follow-up in-depth interviews based in a *TinkerPlots* environment.

The issue of students appreciating the relationship of samples to populations could be the focus of interview-based research if the facility of *TinkerPlots* to select random samples from large data sets is employed during interviews. Using this procedure students' understanding of appropriate sample size and the need to collect repeated samples can be explored. Starting points for such protocols could be based on scenarios initiated with an unexpected sample, such as suggested by Watson (2008). Further, a direct comparison of interpretation of 2-way tables on paper and with the use of equivalent bins in *TinkerPlots* would be possible. These opportunities are currently being piloted by the authors.

6.7 The Final Word

This report has explored many aspects of the use of *TinkerPlots* in an interview setting in order to follow the development of students' understanding of statistics. In terms of the affordances of Anne Watson (2003), the use of *TinkerPlots* software as a basis for student interviews has demonstrated what *is* possible. In today's world of increasing use of technology and especially statistical software, the affordances offered by *TinkerPlots* appear to be valuable assets to both the classroom teacher in developing statistical understanding and the researcher in assessing that development. What *could be* possible depends on the interaction between researchers and teachers, the goals of the curriculum, and the time available. Some suggestions have been made in Sections 6.2, 6.4 and 6.5. It is hoped that feedback in both directions and with teachers becoming teacher-researchers themselves, that the affordances of *TinkerPlots* will be further enhanced in the future.

7. REFERENCES

Australian Education Council. (1991). *A national statement on mathematics for Australian schools.* Melbourne: Author.

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64-83.

Ben-Zvi, D., Gil, E., & Apel, N. (2007). What is hidden beyond the data? Helping young students to reason and argue about some wider universe. In D. Pratt & J. Ainley (Eds.), *Reasoning about Informal Inferential Statistical Reasoning: A collection of current research studies*. Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), University of Warwick, UK, August, 2007.

Burns, R.B. (2000). *Introduction to research methods* (4th ed.). London: SAGE Publications.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-318). Dordrecht: Kluwer.

Chick, H.L. (2007). Teaching and learning by example. In J. Watson & K. Beswick (Eds.), *Mathematics: Essential research, essential practice* (Proceedings of the 30th annual conference of the Mathematics Education Research of Australasia, Vol. 1, pp. 3-21). Adelaide: MERGA.

Chick, H.L., & Watson, J.M. (2001). Data representation and interpretation by primary school students working in groups. *Mathematics Education Research Journal*, *13*, 91-111.

Clements, D.H. (2000). From exercises and tasks to problems and projects: Unique contributions of computers to innovative mathematics education. *Journal of Mathematical Behavior*, *19*, 9-47.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction, 21*(1), 1-78.

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal, 4*(1), 55-82.

Fitzallen, N. (2007). Evaluating data analysis software: The case of *TinkerPlots*. *Australian Primary Mathematics Classroom*, *12*(1), 23-28.

Gibson, J.J. (1977). The theory of affordances. In R. Shaw & J. Bransford, (Eds.), *Perceiving, acting and knowing: Toward an ecological psychology* (pp. 67-82). Hillsdale, NJ: Lawrence Erlbaum.

Ireland, S., & Watson, J. (2008). Concrete to abstract in a grade 5/6 class. In M. Borovcnik (Chair), *TSG 13: Research and development in the teaching and learning of probability*, International Congress on Mathematical Education, Monterrey, Mexico, July, 2008. Available at http://www.ethikkommission-kaernten.at/ICME11/p09_ICME11_TSG13_Ireland_Watson_mb.pdf

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., Harradine, A., & Kazak, S. (2007). Understanding distributions by modelling them. *International Journal of Computers for Mathematical Learning, 12*, 217-230.

Konold, C., & Lehrer, R. (2008). Technology and mathematics education: An essay in honor of Jim Kaput. In L.D. English (Ed.), *Handbook of international research in mathematics education* (2nd ed.) (pp. 49-71). New York: Routledge.

Konold, C., & Miller, C.D. (2005). *TinkerPlots: Dynamic data exploration*. [Computer software] Emeryville, CA: Key Curriculum Press.

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal*, *7*(2), 83-106.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal*, *5*(2), 27-45.

Shaffer, D.W., & Kaput, J.J. (1999). Mathematics and virtual culture: An evolutionary perspective on technology and mathematics education. *Educational Studies in Mathematics, 37*(2), 97-119.

Watson, A. (2003). Opportunities to learn mathematics. In L. Bragg, C. Campbell, G. Herbert, & J. Mousley (Eds.), *Mathematics education research: Innovation, networking, opportunity* (Proceedings of the 26th annual conference of the Mathematics Education Research Group of Australasia, pp. 29-38). Sydney: MERGA.

Watson, J.M. (2001). Longitudinal development of inferential reasoning by school students. *Educational Studies in Mathematics*, *47*, 337-372.

Watson, J.M. (2008). Exploring beginning inference with novice grade 7 students. *Statistics Education Research Journal*, *7*(2), 59-82.

Watson, J. (2008). Eye colour and reaction time: An opportunity for critical statistical reasoning. *Australian Mathematics Teacher, 64*(3), 30-40.

Watson, J.M., & Chick, H.L. (2001a). A matter of perspective: Views of collaborative work in data handling. In M. van den Heuvel-Panhuizen (Ed.), *Proceedings of the 25th conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 407-414). Utrecht: Freudenthal Institute.

Watson, J.M., & Chick, H.L. (2001b). Does help help?: Collaboration during mathematical problem solving. *Hiroshima Journal of Mathematics Education*, *9*, 33-73.

Watson, J.M., Collis, K.F., Callingham, R.A., & Moritz, J.B. (1995). A model for assessing higher order thinking in statistics. *Educational Research and Evaluation*, *1*, 247-275.

Watson, J., & Donne, J. (2008). Building informal inference in grade 7. In M. Goos, R. Brown, & K. Makar (Eds.), *Navigating currents and charting directions* (Proceedings of the 31st annual conference of the Mathematics Education Research Group of Australasia, Brisbane, Vol., 2, pp. 563-571). Adelaide: MERGA.

Watson, J.M., Fitzallen, N.E., Wilson, K.G., & Creed, J.F. (2008). The representational value of hats. *Mathematics Teaching in the Middle School*, *14*, 4-10.

Watson, J.M., & Moritz, J.B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, *37*, 145-168.

Watson, J., & Wright, S. (2008). Building informal inference with TinkerPlots in a measurement context. *Australian Mathematics Teacher, 64*(4), 31-40.

Appendix – Complete Data Set for Data Cards Protocol

Name	Age	Favourite Activity	Eye Colour	Weight (kg)	Fast food meals per week
David Jones	8	TV	Blue	30	7
Brian Wong	9	Football	Green	26	1
John Smith	10	Football	Green	29	0
Adam Henderson	12	Football	Blue	45	5
Andrew Williams	14	TV	Blue	60	10
Peter Cooper	16	Board games	Green	54	2
Scott Williams	17	TV	Blue	65	8
Simon Khan	18	TV	Brown	74	12
Rosemary Black	8	Netball	Brown	24	0
Jennifer Rado	9	Board games	Green	33	4
Anna Smith	11	Board games	Brown	32	1
Kathy Roberts	12	Netball	Brown	32	0
Mary Minski	13	Reading	Green	55	3
Dorothy Myers	15	Swimming	Blue	50	2
Sally Moore	17	Reading	Brown	56	1
Janelle MacDonald	18	Reading	Blue	66	4