REACTION TIME IN GRADE 5: DATA COLLECTION WITHIN THE PRACTICE OF STATISTICS

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

LYN ENGLISH
Queensland University of Technology
l.english@qut.edu.au

ABSTRACT

This study reports on a classroom activity for Grade 5 students investigating their reaction times. The investigation was part of a 3-year research project introducing students to informal inference and giving them experience carrying out the practice of statistics. For this activity the focus within the practice of statistics was on introducing two different ways of collecting data to answer a statistical question, in this case, "What is the typical reaction time of Grade 5 students?" Workbook entries were used to assess students' capacities to engage in the investigation. Results indicated that although the students were proficient with the procedures and measures introduced, they were less able to explain and apply the underlying concepts. The activity provides a suggestion and benchmarks for others wishing to follow student development of concepts related to the practice of statistics.

Keywords: *Data collection; Grade 5; Informal inference; Practice of statistics*

1. INTRODUCTION

Statistics has now been accepted as a strand of the school mathematics curricula of many countries (e.g., Australian Curriculum, Assessment and Reporting Authority [ACARA], 2015a; Common Core State Standards Initiative, 2010; Ministry of Education, 2007). As a newcomer it is still evolving and the statistics education research community is making many recommendations for how the content and pedagogy should emerge. The research is taking place in the context of the capacities of young learners and under the influence of the tertiary theoretical statistics community. What children learn in elementary school must build intuitions that are compatible with what they will encounter later and that are accessible with their current cognitive abilities. The study reported here is a continuation of earlier research based on the practice of statistics in a 3-year longitudinal study with school children in Grades 4 to 6. While reinforcing the practice of statistics as the active participation of students in the core components of a statistical investigation, this study also introduces two methods of collecting data as part of expanding students' experiences, an extension not known to have been reported previously in the literature for elementary school students. The capacity of students to appreciate the two data collection methods, carry out the investigation, and decide which method was "better/more reliable" is assessed based on levels of understanding shown in workbook responses. Transcripts illustrate these with group and class discussion.

Statistics Education Research Journal, 16(1), 262-293, http://iase-web.org/Publications.php?p=SERJ

Underlying the student activity was a continual reinforcement of variation observed, a fundamental concept introduced earlier in the study (English & Watson, 2015a). Acknowledging that many activities for school students provide clean data sets for students and then pose questions arising from the data, this study had students experience collecting two kinds of data and consider which was more appropriate to answer a statistical question.

2. BACKGROUND

2.1. THE PRACTICE OF STATISTICS

The phrase, "the Practice of Statistics" is chosen to describe the activity in which students were involved in this study because at the elementary school level it includes a more straightforward vocabulary. The word Statistics reflects the official curriculum heading of the *Australian Curriculum*: "Statistics and Probability" (ACARA, 2015a). The word Practice reflects the action that takes place: statistics is not a static part of the curriculum. Other phrases used, such as informal statistical inference (ISI), informal inferential reasoning (IIR), and the investigative cycle (PPDAC) are embedded within a wider research context. For elementary students (and some teachers) this vocabulary is too complex for a starting point. There is also historical precedent for the phrase Practice of Statistics, probably first used by Moore and McCabe (1989) in their textbook series that has had many revisions over the years. For them the title expressed their "intent to introduce readers to statistics as it is used in practice. Statistics in practice is concerned with gaining understanding from data; it is focused on problem-solving" (p. xi). The *Guidelines for Assessment and Instruction in Statistics Education* (*GAISE*) of the American Statistical Association (Franklin et al., 2007) for schools followed this intention and used the phrase "statistical problem solving," which is an investigative process that involves four components, each acknowledging the omnipresence of variation: Formulate questions, Collect data, Analyse data, and Interpret results (p. 11). The importance of different aspects of variation is illustrated in the language used: posing the question *anticipates* variability, collecting data *acknowledges* variability, analysing data *takes account* of variability, and interpreting results *allows for* variability. These are the four components of the Practice of Statistics used in the classrooms in this study.

2.2. THEORETICAL BACKGROUND

Carrying out the practice of statistics at the elementary school level requires a type of decision-making that has come to be termed informal statistical inference (Makar & Rubin, 2009). This is in contrast to formal inference (based on theoretical foundations not available to children). Makar and Rubin set the foundation for informal inference that has been the basis for much further developmental research. An informal statistical inference generalises beyond the sample data based on the evidence the data provide, expressing a degree of uncertainty about the conclusion reached. Makar, Bakker, and Ben-Zvi (2011) discuss ISI and then move to considering students' informal inferential *reasoning*, building a theoretical argument based on the early work of Peirce (1931), Dewey (1938), and others. Their foundations of IIR are encased in five categories: statistical knowledge (concepts and ways of thinking), contextual knowledge, norms (collaboration, inquiry) and habits (of mind, of action), inquiry drivers (beliefs, doubt, explanation), and design elements (task, computer tools, scaffolds) (p. 160).

Although developed from different premises, the basis for IIR is not dissimilar to Wild and Pfannkuch's (1999) four-dimensional model of applied statistical investigations, which evolved from their study of their colleagues carrying out statistical investigations. The investigative cycle includes Problem, Plan, Data, Analysis, Conclusion (PPDAC). The complete model of Wild and Pfannkuch (1999) also includes three other dimensions besides the investigative cycle, encompassing types of thinking, an interrogative cycle, and dispositions. Among the types of thinking is consideration of variation. The interrogative cycle parallels the investigative cycle suggesting habits of mind: generate, seek, interpret, criticise, and judge. The dispositions include scepticism, imagination, curiosity, engagement and perseverance. These two theoretical frameworks are basically parallel and employ the tools that are available based on the experience of the learner. The *GAISE* four-step investigative process is focussed on the Practice of Statistics in the classroom and embodies the intensions of both ISI and PPDAC.

The fundamental importance of variation is stressed at every step introduced because without variation there would be no statistics (Moore, 1990). Hence once a problem is posed, the next step is to choose a suitable variable that will provide the required information to address the question. As outlined by Bouma and Ling (2004) the variable must reflect the concept in the question and be measureable in a valid and reliable manner, while at the same time displaying the natural variation that is present through the process of measuring. Lehrer, Kim and Jones (2011) characterize this as a process of modelling variability through the data collected from the measures constructed in order to draw an inference about the problem posed. In asking a question about the typical value in a data set, the focus moves from variation encountered to expectation, perhaps of a single value (Watson, Callingham, & Kelly, 2007), or as Konold and Pollatsek (2002) describe in a metaphor, looking for "signal within noise." Very often the signal or expectation is described as the "average" and is measured with the mean, median, or mode. The application of these concepts to the practice of statistics cannot take place without a context. This requirement was stated forcefully by Rao (1975): "statistics ceases to have meaning if it is not related to any practical problem" (p. 152).

2.3. THE CURRICULUM

The claim that students should experience the practice of statistics from the beginning of their schooling puts pressure on curriculum writers because of the multi-step process involved. In the United States the *Common Core State Standards: Mathematics (CCSSM)* (Common Core State Standards Initiative [CCSSI], 2010) in fact ignores the topic until Grade 6. New Zealand, however, has accepted the challenge and uses the title "Statistical investigation" for one sub-section of the Statistics component of its Mathematics and Statistics curriculum at every level (Ministry of Education, 2007). Over the eight levels of that curriculum Statistical investigation evolves using the investigative cycle of Wild and Pfannkuch (1999). Over the school levels in New Zealand students are expected to be given increasingly sophisticated contexts within which to explore and develop understanding of the stages of the cycle.

The *CCSSM* (CCSSI, 2010) also has eight Standards for Mathematical Practice (pp. 6-8). *The Statistical Education of Teachers (SET)* document of the American Statistical Association (Franklin et al., 2015) translates these specifically for the realms of statistics (pp. 12-17). Paraphrasing these Standards produces the following list relevant to this study.

S1. Make sense of statistical questions posed and persevere in answering them.

S2. Reason abstractly in drawing informal inferences and reason quantitatively during numerical analyses of data.
S3. Construct viable arguments based on evidence from data and critique the reasoning of others.
S4. Model with statistics to find patterns within the variation in data.
S5. Use appropriate statistical tools strategically.
S6. Attend to precision in planning, collecting, and analysing data, and in the statistical language used.
S7. Look forward and make use of structure in data and in the process of the practice of statistics.
S8. Look for and express regularity in repeating the practice of statistics in different contexts.

These Standards of Statistics Practice are reflected in other documents, such as the Australian curriculum's *General Capabilities* (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2013), which includes "critical and creative thinking." In Australia many important tools and procedures are introduced under the sub-strand, "Data representation and interpretation" in the *Australian Curriculum: Mathematics* (ACARA, 2015a). The Australian curriculum also includes four Proficiencies across all content areas: Understanding, Fluency, Problem Solving, and Reasoning.

3. RESEARCH ON THE PRACTICE OF STATISTICS

The complex scenario painted to encompass the authentic practice of statistics helps to explain why some curriculum documents (e.g., ACARA, 2015a) and many research studies have focused in detail on a single component at a time, such as measures of centre (e.g., Mokros & Russell, 1995; Strauss & Bichler, 1988; Watson & Moritz, 2000b), graphical representations (e.g., Friel, Curcio, & Bright, 2001; Lehrer & Schauble, 2004; Watson, 2009), or sampling (e.g., Jacobs, 1999; Rubin, Bruce & Tenney, 1990; Watson & Moritz, 2000a). Few studies attempt to follow the entire statistical investigation from setting the problem to reaching a decision. Lavigne and Lajoie's (2007) Grade 7 students developed their own survey questions to study the research questions they had devised and the researchers analysed the type of reasoning involved at each stage of the investigation. Based on a hypothetical learning trajectory for Grade 6, Meletiou-Mavrotheris and Paparistodemou (2015) began with an assessment of initial understanding and later, after discussion of the results and reinforcement, the students devised their own questions, carried out surveys and analysed the results. English (2014) and English and Watson (2015b) worked with Grade 3 and 4 children, again devising their own survey questions and devising representations to present their conclusions. In all cases surveys, rather than experiments, were the basis for posing questions for studies.

Most studies report posing a question for students or negotiating a question with them (e.g., Thompson, Johnston, & Pfantz, 2006). At this point students may collect the data to answer the question (e.g., Watson & English, 2015b), or they may be given the data from an outside source to be analysed (e.g., Shaughnessy, 2006). The purpose may be the production of a graph or creation of a statistic (e.g., Lehrer et al., 2011; Moritz, 2000), or it may proceed to decision-making with justification provided for the claim made (e.g., Friel, O'Connor, & Mamer, 2006). Having students collect the data themselves to answer the question posed is generally felt to increase the student ownership of the investigation and motivation to complete the analysis providing a reasoned argument in support of the decision made (Van de Walle, 2004).

In their review about reasoning with data, Konold and Higgins (2003) also suggest the four-stage process but then warn …

> Real research, however, seldom proceeds in this orderly fashion. One reason is that conscientious researchers often find themselves backtracking … Experienced researchers look forward from the beginning … They develop and refine their questions and decide what data to collect by thinking ahead to which statistical methods they can use and to the audience they want to convince. Experienced researchers also look backward … And their questions often evolve and change as they discover unanticipated results in the data. (p. 194)

Although recognising that students being introduced to the four stages of the practice of statistics cannot be expected to carry out investigations as experienced researchers, Konold and Higgins want to avoid the recipe-book approach to students' learning. Investigations should be authentic to the extent that the researchers are always conscious of the story that is embedded in the data. At every stage of an investigation, the presence of variation, as recognised by *GAISE*, is likely to cause rethinking of at least part of the investigation, which then requires an acknowledgment of some degree of uncertainty in the decision/s made. Konold and Higgins make this aspect explicit by referring to "backtracking" and linking it to the interdependent phases of research as done by Wild and Pfannkuch (1999).

The combination of Wild and Pfannkuch's (1999) conceptualisation of a statistical investigation as cyclic (with multiple cycles possible), with Franklin et al.'s (2007) constant acknowledgment of variation and Konold and Higgins's (2003) description of possible backtracking, presents a dilemma for the primary classroom. How many of these aspects can children be expected to experience and absorb meaningfully in a single activity? No studies were found introducing different data sets to answer a single question, a situation that might be the result of back-tracking.

The use of software to analyse data has increased the power of statistical enquiry at all levels. At the elementary level the software *TinkerPlots: Dynamic Data Investigation* (Konold & Miller, 2011) provides an environment for students to construct representations and use the associated tools to facilitate learning. This happens while students are exploring plots, perhaps discarding some, and finding one that best tells the story in the data (Konold, 2007; Harradine & Konold, 2006). The benefits of using the software are widely documented at the elementary school level (e.g., Allmond & Makar, 2014; Kazak & Konold, 2010; Lehrer, Kim, & Schauble, 2007; Watson & Fitzallen, 2016). In this study it was used to extend students' opportunities to make decisions about the data collected but that aspect is reported elsewhere.

3.1. THIS STUDY

In the spirit of the practice of statistics, the question for the students to consider in the study reported here was about the "typical" reaction time of students in their grade. Mindful of many years of criticism of students being able to carry out procedures to find the usual statistics for average (mean, median, and mode) but not understanding what they represent (Makar, 2014; Shaughnessy, 2007; Watson, 2007), the word "average" was not used. Makar used an extended inquiry-based approach for Grade 3 students starting with "typical" to develop their own appreciation of the representative nature of average. This approach of using the generic term, "typical" was felt to be sufficiently general to cater for students both with and without previous knowledge of the mean, median, and/or mode.

Acknowledging the complexity of having students carry out every component of the practice of statistics, in this study posing the question was done by the teachers with video clips and hands-on activities to motivate interest in the investigation. From that point, the purpose of the study was for the students to experience the practice of statistics through data collection, data analysis, and decision-making. Keeping in mind the importance of choice of measurement to allow for variation while reliably and validly representing the construct of reaction time (Bouma & Ling, 2004; Lehrer et al., 2011), two contrasting methods of data collection were introduced. All students participated in both methods and were asked later to decide which was more appropriate for the study of reaction time. During and after the data collection, variation was emphasized as was the use of evidence to support a decision. Aware of the tendency of students to believe that mathematics is about truth and *proving* claims (e.g., Chick & Watson, 2002), questions focussed on the certainty/uncertainty with which conclusions were presented (Zieffler & Fry, 2015). The intention was for students to experience and acknowledge uncertainty when interpreting their final results in deciding which method of data collection they thought was "better/more reliable." The activity did not address the subtlety of experience that would take place for professional statisticians; however, the purpose was to make clear to children that different choices were available for data collection when a statistical question is being investigated.

Within the context of developing the practice of statistics in elementary school, taking into account different methods of data collection and the possibility of employing two modes of representation to explore a question of typicality, the following Research Questions were addressed.

Research Question 1: What levels of understanding did the students show in carrying out each stage of the practice of statistics with two different data sets to answer the statistical question?

Research Question 2: What level of consistency was observed between their analyses for the two data sets and why was one chosen as "better/more reliable"?

4. METHODOLOGY

4.1. OVERALL DESIGN

The activity that is the focus of this paper was the fifth of seven major investigations and two shorter lessons that constituted a 3-year longitudinal research project, which began when students were in Grade 4. The activity described here took place at the end of the second year of the project when students were in Grade 5. The aim of the overall project was developing an understanding of beginning inference (Makar & Rubin, 2009) as a basis for statistical literacy. The study was design-based including three cyclic phases: (a) design and preparation of instructional materials for teachers and students, (b) the teaching interventions, and (c) retrospective analyses leading to suggestions for future interventions (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003: Cobb, Jackson, & Munoz, 2016). These phases were implemented for each major activity in the project, with each activity informed by the outcomes of the previous ones. The teachers were involved in preliminary workshops for each activity, gave feedback, and took responsibility for implementation of the activity in their classrooms.

4.2. PREVIOUS ACTIVITIES

Grade 4 began with a benchmarking activity based on students posing multiple-choice questions for their classmates to answer in relation to improvements for the school playground. Responses were collected and represented by students to provide suggestions for the school (English & Watson, 2015b). Due to the fundamental importance of variation to the development of statistical inference (Moore, 1990), the second activity in Grade 4 was based on distinguishing the nature of variation in two contexts: the measurement of one person's arm span by all other members of the class and the measurement of the arm spans of all members of the class once (English & Watson, 2015a). The third activity, also in Grade 4, extended the appreciation of variation to a probability context, where students carried out software simulations of tossing one or two coins to develop theoretical models. The models were confirmed by observing relative frequencies approaching theoretical probabilities as the sample size increased (English & Watson, 2016).

In the second year of the project the first major activity introduced the *GAISE* 4-stage framework for the practice of statistical problem solving (Franklin et al., 2007), as the procedure that leads to making inferences. The context for the activity was a set of five survey questions producing categorical data used to make a decision about the respondents being environmentally friendly. Within the activity, the focus was on the relationship between samples and populations and how this affects the confidence in the inference made (Watson & English, 2015a). The data for this activity were categorical and as part of the activity repeated samples were taken from a known population, with students predicting the proportions of particular responses in the population (Watson & English, 2016b). The activity analysed in this paper was the next in Grade 5 and was devised to reinforce the practice of statistics introduced in the previous activity and extend an appreciation of the data collection phase of the 4-step framework by introducing alternative collection methods and employing numerical data.

The *TinkerPlots* software (Konold & Miller, 2011) was introduced to the students following the benchmarking activity and was used for analysis in the remaining activities. Techniques acquired included using the Plot and associated tools for analysis, using the Sampler to explore probability models with many trials, and using the Sampler to house a population and collect random samples. For the activity analysed here it was felt important to continue to consider the development of students' abilities to create their own displays of numerical data for analysis. Although later in the activity students were provided with their class data in *TinkerPlots* to reinforce and apply the tools there, these data are not analysed in this paper. Of interest, however, was whether their previous experience with the software influenced the hand-drawn plots they produced for reaction time.

4.3. AUSTRALIAN CURRICULUM EXPECTATIONS

Given the context for the investigation involving measurement of length and time, there were seven specific content descriptors across the *Australian Curriculum: Mathematics* (ACARA, 2015a) for Grade 5 that were specifically addressed by the activity.

- Compare, order and represent decimals (ACMNA105)
- Use scaled instruments to measure and compare lengths… (ACMMG084)
- Convert between units of time (ACMMG085)
- Solve simple time problems (ACMMG086)
- Select and trial methods for data collection, including survey questions and recording sheets (ACMSP095)

- Construct suitable data displays, with and without the use of digital technologies, from given or collected data (ACMSP096)
- Evaluate the effectiveness of different displays in illustrating data features including variability (ACMSP097)

Although the four Proficiencies in the Australian curriculum were less extensive than the description in the Mathematical Practices in the *CCSSM* (CCSSI, 2010), of relevance to the Reaction Time Activity for Grade 5 were the following details:

- Understanding – "comparing and ordering … decimals and representing them in various ways;"
- Fluency – "using estimation to check the reasonableness of answers to calculations;"
- Problem solving – "solving authentic problems using … measurements;"
- Reasoning – "interpreting data sets" (p. 31).

4.4. PARTICIPANTS

The participants for this activity were 96 Grade 5 students in four classes in a state-run school in an Australian capital city. The mean age of students was 10 years, 1 month, and 48% were officially classified as having English as a second language (ESL). Only students whose parents gave written permission were included in the study.

4.5.PRELIMINARY LESSON

In preparation for the activity described here based on the practice of statistics, a preliminary activity formally introduced the mean, the mode, the median and the hat plot. The hat plot is a tool in *TinkerPlots* that highlights the middle 50% of the data symbolically with the crown of a hat, while the two brims of the hat distinguish the lowest and highest 25% of the data (Watson, Fitzallen, Wilson, & Creed, 2008). The hat plot is a precursor to the box plot. The lesson demonstrated the three measures of centre using the students themselves, concrete materials, and small numerical data sets. Students also opened *TinkerPlots* and were introduced to the symbols used for the measures in the software, with visual examples of dot plots. The term "average" was used very sparingly throughout to encourage the statistical terminology appropriate for the measures.

4.6. PROCEDURE FOR THE MAJOR ACTIVITY

The activity occupied a full school day in each classroom, approximately 4.5 hours. The context for the activity was human reaction time, a topic appreciated by students not only at their age in terms of their involvement with sport but also in relation to the adult world where safety and speed are issues in many fields of endeavour. The activity began with students playing the "Quick Hands" game, the rules of which are given in Appendix A. Two players each have their own palms together and touch each other's finger tips. They take turns quickly breaking contact to try and touch the back of the other player's hand. After a few minutes of playing, students discussed why one player or the other won, leading to the question, "What is reaction time?" A Smithsonian Institute video clip (http://www.youtube.com/watch?v=t4-MZpqDjVA (2.29mins)) was shown to provide an explanation of reaction time (children were featured in the video) and the definition was linked to the Quick Hands game.

Students were then reminded of the four steps involved in the practice of statistics – 1. Pose question; 2. Collect data; 3. Analyse data; 4. Make a decision (acknowledging uncertainty) – and asked to give examples for each step from the previous "environmentally-friendly" activity where they had made decisions for various populations based on their class sample and random samples from an Australian Bureau of Statistics (ABS) population (Watson & English, 2015b). During this time the word variation was reintroduced and students were asked for examples and definitions. A poster of the four steps was prominent in the classroom during the activity.

Because reaction time can be measured in many different ways, this activity provided the opportunity for students to experience two different methods and two different data sets in order to answer the question, "What is the typical reaction time of Grade 5 students?" The word "typical" was used and discussed by the teacher with the class initially to avoid mention of the three measures of centre introduced in the preliminary lesson. One of the objectives of the activity was to see what tools the students would find useful in their analyses. The question was also a different type from the environmentally-friendly activity where a decision was made in a yes-no context. Here students had the option of selecting either a single typical value or a range of typical values from their numerical data.

Students were asked for and discussed suggestions for measuring reaction time and it was agreed that the Quick Hands game would not provide reliable data. A video was shown of the Batak Pro machine (http://www.youtube.com/watch?v= TWsyyIDwryQ&hd=1&autoplay=1 (1.02mins)), invented to measure and improve peoples' hand-eye coordination, as an example of a high-technology device that was highly reliable. Although impressed, everyone agreed that the class would not realistically have access to such a machine for their investigation.

After some discussion of the question, the first method the students used to measure reaction time was the Ruler Drop. The rules for the data collection using a ruler are found in Appendix B. The teacher and a student demonstrated the procedure and the teacher made sure students realised that the closer the ruler was caught to the bottom (0cm end) as it was held, the faster the reaction time. Students collected the data, recorded them on the class white board, and then wrote the values for the class members in their Workbooks. Students then analysed their data by drawing a representation of the class data set in their Workbooks and answered the following four questions in relation to Step 4 of the practice of statistics.

> Based on your analysis, *What is the typical reaction time for Yr 5 students?*
>
> *Explain* how you reached this conclusion. What tool/s did you use? What process did you follow?
>
> How *certain* are you of your conclusion? Why?
>
> Can you identify any *other issues* with the data collected or the tool/s or process used that might affect your conclusion?

This was the students' second opportunity to create hand-drawn representations of measurement data to show variation in the data collected. In the previous year they had produced representations of arm span data both by hand and using *TinkerPlots* (English & Watson, 2015a).

After a discussion with the students on their results, the second method of data collection was introduced, using the Reaction Timer Test from the ABS *CensusAtSchool* website (http://www.cas.abs.gov.au/tmp_cas/casq_2012_sample.htm Question 13). The instructions for carrying out the test are reproduced in Appendix C from the Student Workbooks. The procedure was based on clicking a mouse after an icon appeared on the computer screen. The result was given in hundredths of a second and values less than 0.1 sec or more than 1.0 sec were omitted to avoid outliers from students trying to anticipate the icon or not clicking the correct spot on the screen. These students repeated the test.

The students answered the same questions in their Workbooks for the analysis based on their hand-drawn representations as for the earlier method. Students were then asked which method of data collections they thought was "better/more reliable" and why?

As part of the design-based research framework, analysis and feedback from teachers after the activity led to recommendations for the following activity.

4.7.DATA ANALYSIS

For the Research Questions, based on students answering the question about the typical reaction time for Grade 5 students, written text responses from the Workbooks were entered into a spread sheet and the representations of the two data sets drawn in the Workbooks were scanned and placed in two individual files for consistency of coding. The questions used in the analysis from the Workbook are presented in Appendix D and the rubrics for coding Workbook responses are found in Appendix E. The items are numbered Q1 to Q11B for reference. The rubrics for all Workbook questions except Q11A were hierarchical, reflecting the sophistication and complexity of statistical understanding displayed in the responses. In coding the process employed by students to suggest the typical reaction time, there were two acceptable possibilities. Students could analyse the data in numerical form, for example lists or tables, or in graphical form, for example value plots or stacked dot plots. The rubrics in Appendix E detail the equivalent levels for the two processes. The responses were coded by an experienced researcher, who consulted with the first author in amending rubrics as was felt necessary. The two agreed on all assigned codes.

5. RESULTS

The results are presented in relation to the Research Questions with examples of students' responses at each level. Because the same questions were asked for the two methods of data collection, the two methods are considered together. Excerpts from class discussions are included to illustrate levels of response and the complexities encountered by the students during their reasoning.

5.1. RESEARCH QUESTION 1

What levels of understanding did the students show in carrying out each stage of the practice of statistics with two different data sets to answer the statistical question?

Answering the first research question required considering the representations created by the students to analyse the two sets of data collected by their classes, their suggested typical values (and reasoning), the level of certainty associated with these values and any issues thought to influence the investigation. Figure 1 contains examples of plots for each of the five code levels for the two data sets. The plot shown for each Code fits the associated descriptor provided in the rubric in Appendix E, with Code 1 responses incomplete or undecipherable and Code 2 unordered or ignoring repeated values. Some students produced lists or tables, whereas others drew graphical forms. Both types of representation could be assessed at Code 3 or above and an example of the most frequently occurring type is shown in Figure 1, which was an unordered value plot or attempted calculation. Code 4 responses produced ordered data showing frequency, but only at Code 5 were typical values also shown on the plot. The percentages summarise all plots at that code level. Over half of the students produced a stacked dot plot or equivalent for each type of data but fewer went on to indicate the typical value/s on the plot.

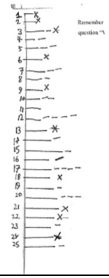
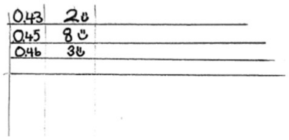
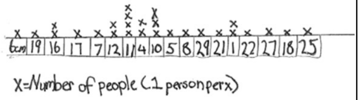
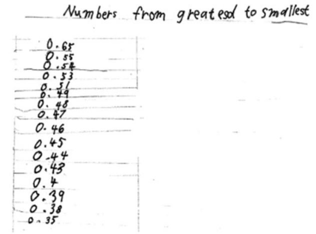
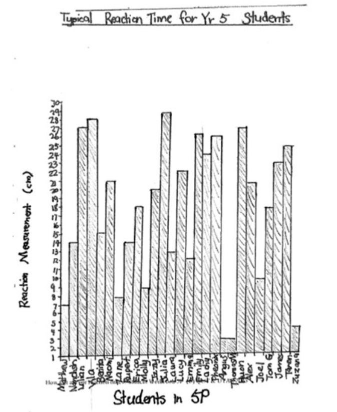
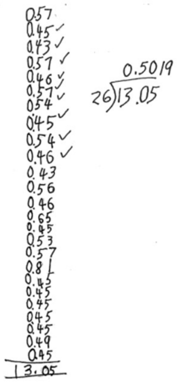
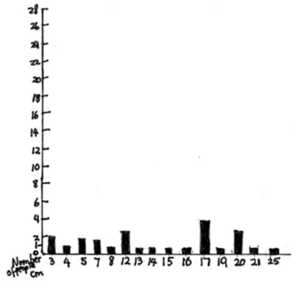
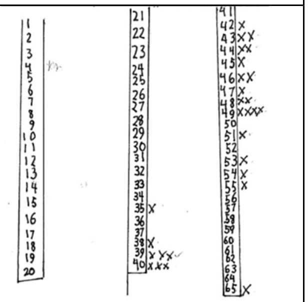
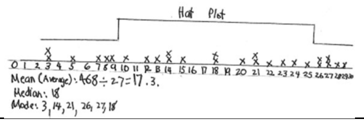
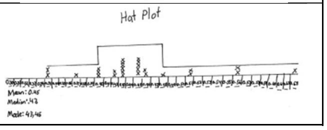
	Ruler Drop		ABS Timer	
Code	%	Plot for Ruler data	%	Plot for Timer data
1	5%		8%	
2	9%		15%	
3	22%		8%	
4	48%		53%	
5	16%		15%	

Figure 1. Examples of students' hand-drawn plots for the two data sets at each code level, with the percentage occurring for each Code

The success in suggesting “typical” measurements for the ruler drop and time from the ABS timer is reported in Table 1. Overall students found this value quite easy to estimate.

Table 1. Levels of response for suggesting the typical measurement for Q2 and Q7

Code	Description of typical measurement given	% Ruler (Q2)	% Timer (Q7)
0	Number larger than 30 (length of ruler) or outside of range for timer (0.1 < value < 1.0)	6%	4%
1	Number within the range of values in the plot but not near centre of data	11%	4%
2	Number “near” middle of plot or near often repeated value (mode)	82%	92%

The results for Workbook questions Q3 and Q8, Q4 and Q9, and Q5 and Q10 (Appendix D) are found in Tables 2, 4, and 5, respectively, showing the percentage of each code level for the two data sets along with examples of responses. It should be remembered, however, that all student analysis and discussion about the ruler drop data took place before the analysis and discussion of the ABS timer data. As can be seen glancing through the example explanations, they are similar across methods at the various code levels.

For questions Q3 and Q8, the rubric reflected both the appropriateness of the typical value suggested and whether the explanation was consistent with this value. Following their success in predicting (Table 2), for both methods about two-thirds of students could justify their choices appropriately. At lower levels explanations were either inconsistent with the value chosen (Code 2) or vague without specific reference to a procedure (Code 1). For each method about 10% did not respond or address the question (Code 0).

Of interest are the conversations among students as they made decisions. Although working in pairs around the classroom, all students decided their own typical values. The exchange that follows is related to the class data in Figure 2 and illustrates some of the dilemmas encountered by the students. Students 48 and 50 formed a pair working together. As can be seen in the exchanges, it is difficult to avoid reference to “average” and it can be confusing. Although the students knew the definition of an outlier as an extreme value, being able to decide *how* extreme an outlier had to be was more difficult. In the classroom, the responses were valued based on the justifications provided and there was no “correct” answer.

Table 2. Reasons for explaining how the typical value was chosen (tool/s and process) for Q3 and Q8

		Ruler Drop		ABS Timer
Code	%	Tools (Q3)	%	Tools (Q8)
0	12%	The tool we used was a ruler the[n] we drooped [dropped] it and looked at the number below the thumb.	10%	[No response]
1	14%	I used a bar graph. Because I added them all up, and checked.	8%	I collected data, put them in a chart and made a graph. I reached my conclusion by putting the seconds in down below then up the top I put in the data we collected.
2	7%	Inconsistent with data: Added up the totals to 344 and ÷ by 27. Because it was the 'mode' of this graph. 28 ÷ 2 = 14	14%	I took 0.42, 0.43, 0.46 and 0.49 and then average them. I arranged the number[s] an[d] I got 0.43. 0.81 ÷ 2 = 0.405
3	64%	I used the mean to find this answer. I used the mode, and figured the mode out because it occurred the most. We looked at the cluster and decided on its range. We then estimated the middle of the range.	68%	We used the mean tool to find our conclusion which was 46 sec. I kept crossing each dot on my plot out until there is only one left. Median. Because there was more people that got that result.

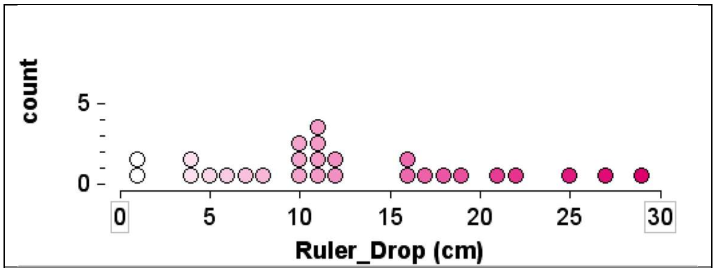


Figure 2. Class data from the Ruler Drop data collection

Student 50: If you're writing typical, does it mean the average?
S48: Yeah.
S50: Did you get 11 as the average?
S48: Yes.
S50: Did you use the calculator?
S48: No, the typical means the average, but it also has to come from how many people got the highest number [mode] I guess.
S50: By typical does it mean the average or the mode? [Question to Teacher.]

Teacher indicates either is okay if explained.

S48 continues to add up all values from his table.

S50: And now you have to divide it by [counts number of students in table], how many things are there, 335 [344].
S48: Just tell me how many people there are, that’s all I want to know.
S50: [counts students again] 27.
S48: Divide by 27. [does that on his calculator] 13? … so the average is 13 …
S48: S77 [sitting nearby]? What did you put as your typical?
S77: 11.
S50: I put 13 …
S48: The mean we got 13cm which is the average, and we were going to put that down but then we realised no-one had 13cm [on the plot] so it’s just an outlier that caused it.
S50: There’s no outliers.
S48: Yeah, 29.
S77: 27 would be the outlier.
S48: And 29.
S50: But that’s not really an outlier.
S48: Yeah but still 29 and 27 are a lot, lot different to these bottom ones.
Teacher: So is 1, isn’t it? (laughs).
S50: There’s two 1s.
S48: Yeah but that’s not so far away as [the others].

The final Workbook responses for these three students are shown in Table 3, illustrating that the final conclusions drawn from the data and discussion were not the same for the two students working together.

Table 3. Workbook responses from the students in the conversation

Student	Response Q2	Response Q3
S48	11 cm is the mode. 13 cm is the mean.	I used a plot graph to show which people had fast reactions and which people don't. However I think the typical is 11 cm because no-one had 13 cm, which was caused by outliers.
S50	13 cm is the average time for using the mean I used I got 13 cm. The mode was 11 cm.	The thing I used was the mean and I added all the numbers and divied [divided] them. I used mean because it is the most reliable here because there are hardly any outliers.
S77	The typical reaction time for year five students is 11 cm.	I used my graph to get to my conclusion and mode too.

The reasoning that students used to determine their typical values for the ABS timer data is illustrated in the classroom discussion held in one class after the students had filled in their Workbooks. The class data they were referring to are shown in Figure 3. The first student the teacher asked said all of the numbers 0.39, 0.42, 0.43, 0.46, and 0.49 were typical. When asked why, the student said because each happened three times. The teacher (T) then asked for other comments, the responses displaying a range of methods and a few misunderstandings. Similar to the previous extract the teacher did not make judgments on the choices of method to find the typical reaction time and used another student to point out the error in deciding the total number of data values. The extract shows some students comparing various conventional methods of determining central tendency, whereas others invented unconventional methods such as finding the mean of the modes or rounding values and making another stacked dot plot. Reinforcing the proper pronunciations of decimal numbers happened repeatedly throughout all classroom discussions.

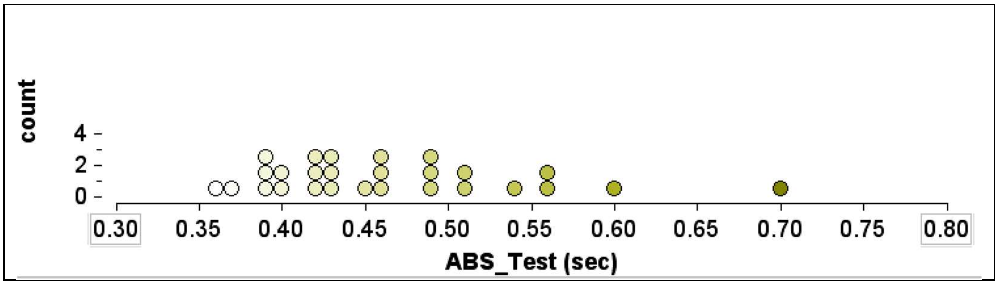


Figure 3. Class data from the ABS data collection

S55: Well I thought that at the beginning but I thought like a bit too many typical numbers, so I did it the mean way.

T: So you thought that at the beginning and then you went and used a different tool?

S55: Yeah.

T: … and how did you do that?

S55: I added up all of the …

T: Every single value?

S55: Um, yeah and then I divided it by … 14.

T: 14? Why 14?

S55: Because there were 14 values.

T: 14 values? Were there?

S48: She means, I think she means different values but that's not …

T: … so we should have divided by what?

S48: 27.

T: 27. Okay, did anybody do that? S48?

S48: Um I did it that way [modes] … And then I realised that … if I did the mean maybe it would be, narrow it down …

T: What was the answer?

S48: "forty-six."

T: "Decimal four six."

S48: Yeah.

T: Quite a difference between "forty-six" and "decimal four six."

S48: Just a little bit … [general laughter]

T: … S15, what did you do?

S15: I did the median.

T: … So you put them all … In order? … Good, ascending order.

S15: And I, I counted like inwards but as you can see like some of them had three times so sometimes you had to stay on them and then count over the next one and …

T: Mmm, jump back and forth …

S15: … But I got "zero point four five" as the median.

T: Right, so the mean we got .... um, "decimal four six," and for the median you got "decimal four five."

S15: Yep.

T: Okay, so little bit of difference there, okay, right, … who had something completely different from any of that?

S08: Well I did something diff, a bit different.

T: Good.

S08: I used the mode but in a different way, I didn't just use all four of them, I used,

S08 had missed one of the five modes, which is discussed, then S08 continues.

S08: I um, I added them all, I averaged them.

T: So you averaged the modes?

S08: 42, 43, 46 and 49, I averaged those and it came to "zero point four five." [for five modes would have been 0.438]

T: Okay, interesting, so that's the average of the modes.

S08: Four of them.

T: Which gave us the same result noticeably as the median, didn't it? Um, S48?

S48: I did mean, median and mode and then I saw which one was most accurate.

T: … do you want to expand on that for us?

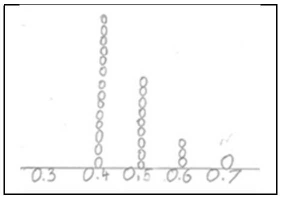
S48: … I just sort of guessed so, yeah … I used the mean.

T: You decided on the mean, alright, what did you do S92?

S92: I did, well I rounded off all the results and I got, so I got ...

T: For every person's results, you rounded?

S92: Yes, so I drew a dot plot so and um, the numbers I got were point 3, point 4, point 5, point 6, and point 7 and then I rounded off all the results in the table and the typical number I got was point 4.



[from Workbook]

For Q4 and Q9, asking for the students' confidence in their answers, a relatively even distribution of responses resulted for both data collection methods. At Code 0, no actual reason was expressed for the confidence, whereas at Code 1 the reason non-specific. Reliance on the skill of calculating was expressed at Code 2 and the meaning associated with the tool used was included at Code 3. Finally at Code 4, uncertainty was expressed due to the nature of the sample.

Table 4. Degree of confidence associated with the typical reaction time value chosen for Q4 and Q9

		Ruler Drop		ABS Timer
Code	%	Confidence (Q4)	%	Confidence (Q9)
0	22%	I am 75% certain of my conclusion because I think that my conclusion would turn out well. I am not very certain because my friends said [what] I had recorded was wrong.	17%	Certain, I do not know why. I'm 50% sure.
1	14%	I am certain because the numbers are very close to each other and I double checked. I am fairly certain because after listening to my peers we all reached the same answer.	15%	Yes, I am confident because I spent a lot of time working out. I am certain of this conclusion because it is a reliable source.
2	26%	I'm not that certain because when I used other tools I got different answers. I am very sure my answer is correct because I went through it 5 times and it was still 11.	20%	I am certain because I've checked. I'm certain of my conclusion because I think I followed my process and did well.
3	22%	I am more than half certain because under the crown the numbers were ten to twenty five so inbetween ten and twenty five was around seventeen and eighteen. [for value of 27] I'm very certain because most pepole [people] got 20 to 29.	26%	I am 100% certain because there is much, much more 0.45 then any other else. A bit certain because the median might be the average but I'm not certain. I am uncertain because the average is 0.58 but most people got 0.46.
4	17%	I am not very certain because we only recorded the reaction time for one year five class. I am 75% sure because this is only 1 class we have tested on so maybe if we do more classes the conclusion may change.	23%	I am not very certain because only year 5G students did it and not all year 5 students. I am not very certain because we only surveyed 27, year 5 students out of 280,000.

Finally students were asked what other issues they could identify related to the investigation they had undertaken (Q5 and Q10). This question was difficult for many students with nearly half of the students unable to be critical in any way about the investigation they had undertaken (Code 0). At Code 1 about a third of students suggested various potential errors in the data collection or analysis, such as practicing to increase skill or making errors in calculations. Around 10% of responses suggested perhaps the wrong tool was used for analysis (Code 2) and slightly fewer responses reflected sampling issues as potentially troubling (Code 3).

Table 5. Issues that arose associated with the investigation for Q5 and Q10

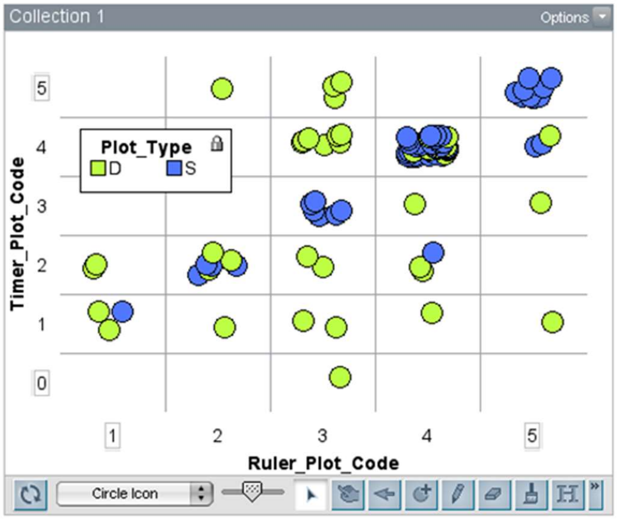
		Ruler Drop		ABS Timer
Code	%	Other issues (Q5)	%	Other issues (Q10)
0	45%	We didn't find any issues. Most of the numbers were 2 or 1. I am positiv[e]ly sure there is nothing wrong with my graph.	44%	I thought that it was a good tool to identify the reaction time. I don't think there are other issues. No, I do not have any affect [sic] on my conclusion because the task was easy to understand.
1	33%	Some issues that effected our conclusion was that people didn't always get a whole number. Yes if the missing student (Nolan) came back and got 30 (outlier) there would be a huge change in the results. An issue of the data collected is that the measurement might not be precise.	36%	The data may be wrong. You could cheat the system with lots of pra[c]tice the person could have lied or the computer could be wrong. There were 2 out leirs [outliers] this time. So that migh[t] [be] a problem.
2	12%	The tools that we use is [sic] median. Another reason that my conclusion may not be correct is that 17.33 is only the mean not the median or the mode.	10%	You could use a bar graph. The mode might be better to use because it will represent more students to get a typical answer.
3	9%	Each class might have different results. And some of the students may of not performed the experiment effecting our results. Yes it could be the whole world for example. If we did it more than once some people might improve or got worse.	9%	If we recorded all the year five classrooms and created a new average would be better. If we did it more than once it might affect the conclusion. Mabey [maybe] if we had a diff[e]rent class because you had a diff[e]rent class and they might have people with diff[e]rent skills than this class.

Progressing through the analysis of their data to determine typical reaction time (Q2 and Q7), 77% of students showed the capacity to predict adequately for their data using both methods, whereas a further 19% reached the highest level on one of the methods. Students continued to give appropriate explanations (Q3 and Q8), with 56% achieving Code 3 on both methods and 21% doing so on one of the methods. Expressing confidence (Q4 and Q9) and appreciating the issues involved (Q5 and Q10) in the analysis were more difficult for students, with only 48% responding at Code 3 or higher for confidence for at least one method and 30% responding at Code 2 or higher for issues for at least one method.

5.2. RESEARCH QUESTION 2

What level of consistency was observed between their analyses for the two data sets and why was one chosen as "better/more reliable"?

Research Question 2, addressing the association between the analyses for the plots created for the two data sets, was considered in relation to whether the students produced the same or different types of representations for the two data sets. Fifty-five percent of students produced the same type of representation for both data sets whereas 42% produced different types. Figure 4 shows the association between the codes for the two types of representations. The indicative correlation between the two sets of codes was 0.619. There was more variation between the codes for those who used different representations, even though the mean scores on each were very similar. This is seen in Figure 5 where the differences of the two codes are plotted for all students. For neither group of students was there an improvement in the placement of the mean in drawing the second representation from drawing the first.



D=different type of plot for the two data sets
S=same type of plot for the two data sets

Figure 4. The association of code levels for the two plots

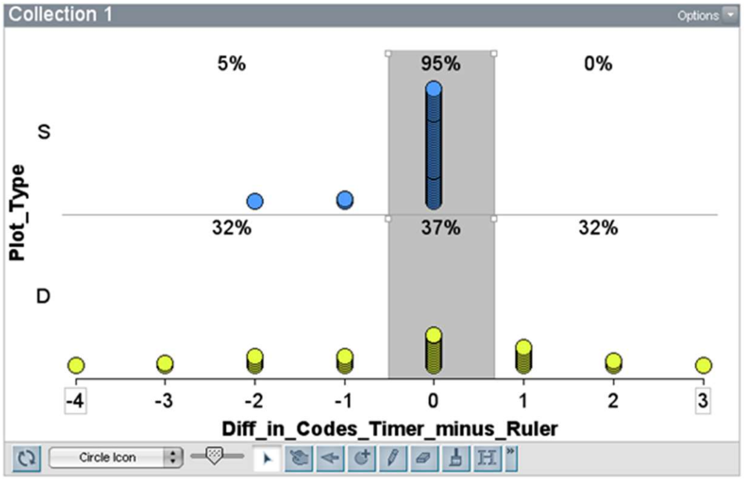


Figure 5. *Difference in code by Plot Type: S = Same, D = Different, Mean (S) = -0.073, Mean (D) = -0.171*

The correlations across the responses were also similar for the three pairs of questions, being 0.600 for Q3 and Q8, 0.531 for Q4 and Q9, and 0.515 for Q5 and Q10.

The correlation of the code levels for the individual questions is reflected in the indicative correlation for the sums of codes for the five questions (Q1 to Q5) and the second five (Q6 to Q10), which was 0.601. This leads to an expectation of little difference for the sums of codes for the two methods. This is shown in Figure 6, where although there is a shift away from the lower total scores, the median was 11 for both sets and the means were not significantly different.

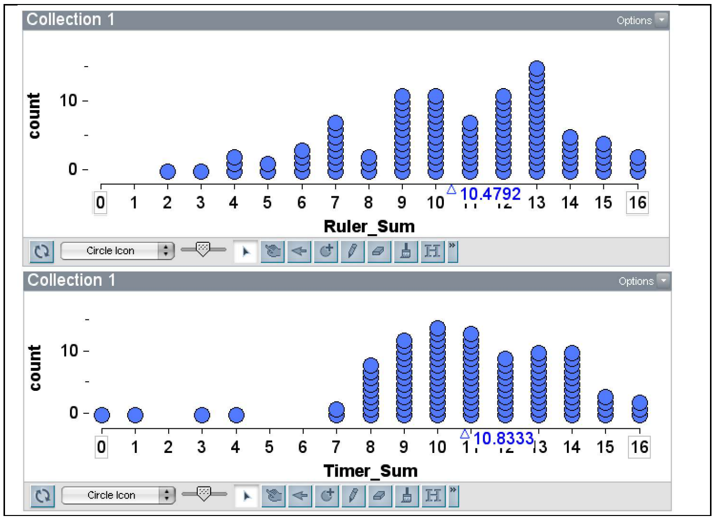


Figure 6. *Total scores for the two data collection methods for Research Question 1*

Workbook question Q11 was initially coded in terms of the choices made by students as to which method was "better/more reliable" in characterising typical reaction time. The results are shown in Table 6.

Table 6. Which method is better/more reliable? (Q11A)

Method	Percentage
No response	17%
Computer	71%
Ruler	8%
Same	4%

As there was no “correct” answer to question Q11B, the coding for the question was based on the sophistication of justifications given for the choices based on the appreciation of the need for accuracy and the fact that the ABS actually measured time rather than centimetres that would require translation into seconds. The results are shown in Table 7 with example responses. At Code 0 about a quarter of responses did not provide an interpretable response. At Code 1 a method was chosen because it was more fun or more difficult. About half of responses made a decision based on accuracy, often related to the ease of making the measurement (Code 2). Because of the nature of the observations and their different units of measurement Code 3 responses recognised that the ABS timer was actually measuring reaction time whereas the ruler measurement was in centimetres and would need to be translated somehow into time, which they did not have the resources to do.

Table 7. Explanations for the choice of the better/more reliable method (Q11B)

Code	%	Example responses
0	26%	Definty [definitely] the computer one. Out of the 2 methods the ruler measurement is way better by 22. I think that both of these two methods have an equal amount of determining the reaction time, because they both test the reaction time a lot. The one on the computer is better because it’s a computer.
1	8%	Ruler is better because it is more easy to make the graph and its fun. I think when we had to react to the falling ruler was better because there was a count down. I think the one with the ruler is better because some people aren't good with laptops and computers. Rulers, because the mouse moves to[o] much and is harder for most people (like me) and the mouse kept moving to one side. Both of them are good at bits for example the ruler tells how you can do it but the computer shows how fast. The second because it's more fun.
2	51%	The computer becaus[e] its ac[c]urate. I think the second method is the better because its an easier way to get the reaction time. I think second one is better because it is easy to collect the data. The ABS test (seconds) is better because it is more accurate than the ruler drop[p]ing because it’s newer tecnoliy [technology].
3	14%	The ABS Test because it times you. The rular [ruler] one measures in cm. The computer is a better method because it would be more accurate and gives the correct unit of measurement. I think the computer one is better because it times you at the end.

The class discussion after students had completed their Workbook entries generally reflected the entries. An interesting interpretation, however, was suggested by one student verbally to his class.

S24: That computer, anyone can do it but the other one's [ruler's] harder. It's a bit like catching a ball, like someone's throwing a ball hard at you, you have to quickly get it. But that one's [computer's] like, you just have to wait until it comes up and then push a button.

Another exchange occurred among several students and the teacher at the end of collecting the ruler drop data.

S102: Um, we um, we may have a few issues because when we collected the data, there may have been a margin of error.

T: A margin of error, that's a very nice word, margin of error, and what is your margin of error, would you like to explain to the class what you mean by margin of error.

S102: It's like, it's like how many centimetres off it can be from the actual measurement.

T: So you're saying, there might be a chance, and you're very right, that when you're catching your ruler, you may have measured it a little bit differently.

The teacher expanded her comment by giving an example from collecting the measurement from another student. This student then explained what he thought caused the margin of error.

S42: Like before you like grab it you need to make like a gap like (indicates with his fingers how they were instructed to hold their hand in preparation for grabbing the ruler when released), and then if the gap is bigger you will like, your reaction time will be slow but if your gap is like just a bit more closer then you [will be faster].

Another issue was suggested by a student related to the positioning of the ruler with respect to the hand before dropping.

S60: It kind of depends, like if you're like holding the ruler *in* the hand or *above* the hand, you'd want to hold it above the hand because if you put it like in the hand (demonstrating distance between ruler and catching hand prior to release), you might drop it, and the hand will like get 7 or zero.

Although issues of variation and uncertainty were implicit in many of the responses justifying which method was better/more reliable, the terms were not used explicitly in the responses. Given the classroom observation of the students collecting their data, it is not surprising that the ABS time was generally chosen as better/more reliable.

6. DISCUSSION

6.1. STUDENT CAPACITY

The first focus for the Discussion is Research Question 1, related to the capacity of students to develop further their understanding of the practice of statistics and to absorb issues related to using two different data sets to answer the same statistical question. This was the students' second major activity in which the four steps of a statistical investigation (Franklin et al., 2007) provided the framework for the class investigation. The four steps, displayed on each classroom wall, were emphasised throughout by the teachers and the Workbook structure.

The researchers considered the students' analyses for the two data sets together because the questions were the same and the students completed them in the same order. The activity was set up in this manner to encourage students to think about the influence of the two data sets. It also provided reinforcement for the overall investigation steps that engage students in the practice of statistics. Students were very careful with both types of data collection. Very few of the ruler drops or timer runs had to be repeated but these cases were good situations for students to experience in terms of realistic data collection. About two-thirds of students could produce a meaningful hand-drawn representation of their data. Although usually not indicating the typical value on their representations, most could suggest a reasonable value when asked and again about two-thirds could give a meaningful explanation of their method of deciding the value. Students were less successful with the more sophisticated task of expressing reasons for limited confidence in their conclusions holding for a wider population of Grade 5 students.

With respect to Research Question 2, the levels of understanding displayed were similar for the two analyses (cf. Tables 1, 2, 4 to 6). Although the overall means for the representations drawn were similar, Figure 5 shows that inconsistencies occurred mainly for those who chose different types of plots for the two methods. Because there was not a great deal of time for discussion between the two data collections, it is perhaps not surprising that there was little improvement for the mean total scores on the five questions for the two methods (cf. Figure 6).

The exposure to the two data sets was intended to add interest and awareness of issues associated with data collection. It was not felt appropriate, however, to introduce formally the more complex notion of "backtracking" (Konold & Higgins, 2003). In fact there was implicit backtracking in the Workbook task to represent the data collected under each method when the instructions said, "Remember to ensure that your representation helps you to answer our question, 'What is the typical reaction time for Yr 5 students?'" Evidence of erasures on the scans of the plots/tables produced show that some students did go back and rethink their initial representation. Thinking about which data collection method was "better/more reliable," however, was the major extent of the class discussion. Nearly two-thirds of students could justify their choices in meaningful ways for their ages. The intention was to refer back to this activity in later ones to remind students that there may be various possibilities for data collection.

The question that asked about issues arising during their investigations (cf. Table 5) was intended to encourage critical thinking about the procedures the students had used, reflecting the aims of the "critical and creative thinking" general capability in the *Australian Curriculum* (ACARA, 2013) and of the Partnership for 21st Century Skills (2009). This might have been a point leading to backtracking for some students but generally these Grade 5 students did not take on the challenge. Asking students to consider issues in investigations that could affect their confidence in the conclusion was another aspect of the practice of statistics to be reinforced in later activities. As well this strategy was meant to reinforce the appreciation that informal inference (Makar & Rubin, 2009) as carried out through the practice of statistics produces evidence to support a conclusion but is not a proof.

In the spirit of the design-based methodology employed for the project (Cobb et al., 2003, 2016), there was discussion with the teachers that led to suggestions for future interventions and the reinforcing of the interpretation and plotting of decimal numbers less than one. Among these suggestions were the need for helping students express in meaningful language the process of finding the mean, median, and mode, as well as recognising uncertainty in their final conclusions. The next activity for students was planned to reinforce these concepts and others involved in carrying out the practice of statistics (Watson & English, 2016a).

Of some interest was the previous exposure to *TinkerPlots* and how this influenced the hand-drawn representations created in the Reaction Time Activity compared to those drawn showing arm span before using *TinkerPlots* to do so (English & Watson, 2015a). Although 20% students had in the previous activity created meaningful ordered frequency plots with grids or icons, which were equivalent to stacked dot plots as created in *TinkerPlots*, none actually employed the dot plot format with inscriptions (x's) as seen in Figure 1. For reaction time data, 42 students (44%) created stacked dot plots as their representations for the ruler drop and 45 students (47%) for the ABS timer, although not all were in the top two code levels. There was also less variation across the types of representations produced for reaction times than for arm span measurements.

6.2. THE STANDARDS OF STATISTICAL PRACTICE AND THEIR IMPLEMENTATION

Besides knowing the tools and terminology as outlined in the content descriptors of curriculum documents, there was evidence of students taking on the Standards of Statistical Practice as translated by Franklin et al. (2015). Students made sense of the reaction time context with the question of "typical time" and persevered in both methods of data collection to describe their typical time (Standard 1). Most students reasoned abstractly from the quantitative data they collected and analysed (S2), making viable arguments to support their typical reaction times (S3). The modelling at this level resulted in the suggested "typical values" (and variation was acknowledged) (S4). Generally students demonstrated they could use the tools (mean, median, mode, hat plot) introduced in the preliminary lesson (S5). The students were very precise in carrying out the rules for data collection, particularly for the ruler drop, and in producing their representations, explaining carefully in their Workbooks with statistical language (S6). They followed the structure of the practice of statistics and looked for the structure in the data to suggest the typical value (S7). Finally in this activity the students had the opportunity to undertake the practice of statistics across two contexts for data collection (S8). Hence, not only did the Reaction Time Activity satisfy the content and proficiency descriptors of the *Australian Curriculum: Mathematics* (ACARA, 2015a) but also it provided examples of the Standards set by the American Statistical Association (Franklin et al., 2015) in parallel with the *CCSSM* (CCSSI, 2010).

The implementation of the activity for Grade 5, however, presented students a four-step framework based on *GAISE* (Franklin et al., 2007). Having a poster on the wall (see e.g., Watson & English, 2015a) was felt appropriate at this grade level, in order to reinforce the investigation process. The focus on informal inference (Makar & Rubin, 2009), which is so essential at the school level to build intuitions for later study, was included in the activity through asking questions on how confident students were of their conclusions and what other issues might influence their decisions (see Tables 4 and 5; Zieffler & Fry, 2015). The wording "informal inference" was not used at this grade level but acknowledging variation in the data and questioning certainty in decisions were included often by all teachers in the class discussions.

6.3. LIMITATIONS

Following the preliminary lesson on measures of average there could have been a pre-test on students' memory for the three terms at the beginning of the activity. The researchers, however, wanted to observe what the students remembered and used naturally without prompting. Although the three measures received equal attention during the formal lesson, in explaining how they found the typical value for the ruler drop, of the 96 students only 12 students used the word mean, 6 used median, and 27 used mode. For the ABS timer, only 16 used the word mean, 12 used median, and 25 used mode. Some used more than one and others described a process without naming it. Overall, for the ruler drop 45% of students used at least one word, whereas 49% did so for the ABS timer. The authors speculate that given the visual presentations, the mode was the easiest to recognize and describe. No measure was given preference in the coding of responses.

Although from a technical design perspective it might have been desirable to alternate data collection methods in different classes, all students completed the two methods in the same order, using the ruler first. It is likely that the discussion of the first results influenced choices made in employing the second method, the ABS timer. Classroom organisation based on the teacher feedback and the time involved for the two methods meant it was not possible to counterbalance the order in which classes began the activity.

The transferability is limited to some extent by the practical aspects of the implementation in the classroom and the particular framework employed as the practice of statistics. As pointed out in the introductory sections, however, the relevant aspects of the foundations of Makar et al. (2011) and Wild and Pfannkuch (1999), as set out by *GAISE* (Franklin et al., 2007), would appear to be sufficiently general to be consistent with the outcomes of other studies.

6.4. EXTENSION

Although it would have been possible to translate the cm measurement from the ruler drop into seconds using the formula $d = \frac{1}{2}gt^2$ and solving for t, where d is the distance of the drop, $g = .980\text{cm/sec}^2$ is the acceleration of gravity and t is time, the resulting equation of $t = \sqrt{\frac{d\text{cm}}{(490\text{cm/sec}^2)}}$ involves a square root, which is not a procedure that is in the *Australian Curriculum: Mathematics* (ACARA, 2015a) for Year 5. Both gravity in the *Australian Curriculum: Science* (ACARA, 2015b) and square root in the Mathematics curriculum are introduced in Year 7. This would be a good extension for high school students who would be likely to find that the typical reaction time using the ruler drop would be less than for the ABS timer. This would lead to discussion about the way in which reaction time is measured and its purpose. Which method provides quicker time values may not be an issue because most likely the reason for measuring reaction time is comparison, either with others or with oneself to look for improvement. It was also felt important for students to experience different ways of measuring the same construct.

6.5. CONCLUSION

The students in this study were reinforcing their previous encounter (Watson & English, 2015b) with the implementing the practice of statistics (Franklin et al., 2007) in a meaningful context as required by Rao (1975), encountering variation (Moore, 1990) within each data collection method and informally between the two methods. Students were asked for their expectation of the “typical” reaction time amongst the variation in the data collected (Konold & Pollatsek, 2002; Watson et al., 2007), and to make an evidence-based decision about their confidence in that expectation (Makar & Rubin, 2009). Finally they were asked to solve a problem (Franklin, et al.; Moore & McCabe, 1989) concerning which method of data collection was more reliable. The extracts from class discussion and Workbooks suggesting unconventional methods for finding typical reaction times illustrated the value of focusing on data representations to determine typical times rather than just being taught to apply conventional algorithms to lists of data values. Although more reinforcement is needed in the future, these students have made a reasonable start and most appear to have the capacity to develop realistic intuitions about the practice of statistics across the school years.

ACKNOWLEDGEMENTS

This article arose from research funded by Australian Research Council (ARC) Discovery Grant DP20100158. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the ARC. We wish to acknowledge the excellent support of our senior research assistants, Jo Macri and Dr Ben Kelly.

REFERENCES

Allmond, S., & Makar, K. (2014). From hat plots to box plots in *TinkerPlots*: Supporting students to write conclusions which account for variability in data. In K. Makar, B. deSousa, & R. Gould (Eds.), *Sustainability in statistics education* (Proceedings of the 9th International Conference on the Teaching of Statistics, Flagstaff, Arizona, July 13-18). Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_2E1_ALLMOND.pdf

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2013). *General capabilities in the Australian Curriculum, January, 2013 (updated September 2014).* Sydney, NSW: ACARA.

Australian Curriculum, Assessment and Reporting Authority. (2015a). *Australian Curriculum: Mathematics, Version 7.4, 30 March 2015.* Sydney, NSW: ACARA.

Australian Curriculum, Assessment and Reporting Authority. (2015b). *Australian Curriculum: Science, Version 7.4, 30 March 2015.* Sydney, NSW: ACARA.

Bouma, G.D., & Ling, R. (2004). *The research process*. (5th ed.). South Melbourne, VIC: Oxford University Press.

Chick, H.L., & Watson, J.M. (2002). Collaborative influences on emergent statistical thinking – A case study. *Journal of Mathematical Behavior*, *21*, 371-400.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13.

Cobb, P., Jackson, K., & Munoz, C. (2016). Design research: A critical analysis. In L. D. English & D. Kirshner (Eds.), *Handbook of international research in mathematics education* (3rd ed.) (pp. 481-503). New York: Routledge.

Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics*. Washington, DC: National Governors Association for Best Practices and the Council of Chief State School Officers. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Dewey, J. (1938). *Logic: The theory of inquiry.* New York: Henry Holt & Company.

English, L.D. (2014). Promoting statistical literacy through data modelling in the early school years. In E. Chernoff & B. Sriraman (Eds.), *Probabilistic thinking: Presenting plural perspectives* (pp. 441-457). Dordrecht: Springer.

English, L., & Watson, J. (2015a). Exploring variation in measurement as a foundation for statistical thinking in the elementary school. *International Journal of STEM Education*, *2*(3). DOI 10.1186/s40594-015-0016-x

English, L.D., & Watson, J.M. (2015b). Statistical literacy in the elementary school: Opportunities for problem posing. In F. Singer, N. Ellerton, & J. Cai (Eds.), *Problem posing: From research to effective practice* (pp. 241-256). Dordrecht: Springer.

English, L., & Watson, J. (2016). Development of probabilistic understanding in fourth grade. *Journal for Research in Mathematics Education*, *47*, 27-61.

Franklin, C., Bargagliotti, A., Case, C., Kader, G., Scheaffer, R., & Spangler, D. (2015). *The statistical education of teachers*. Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/education/SET/SET.pdf

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A preK-12 curriculum framework.* Alexandria, VA: American Statistical Association. Retrieved from http://www.amstat.org/education/gaise/

Friel, S.N., Curcio, F.R., & Bright, G.W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*, 124-158.

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* (pp. 117-137). Reston, VA: National Council of Teachers of Mathematics.

Harradine, A., & Konold, C. (2006). How representational medium affects the data displays students make. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education* (Proceedings of the 7th International Conference on Teaching of Statistics, Salvador, Bahai, Brazil. Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots7/7C4_HARR.pdf

Jacobs, V.R. (1999). How do students think about statistical sampling before instruction? *Mathematics in the Middle School, 5*(4), 240-263.

Kazak, S., & Konold, C. (2010). Development of ideas in data and chance through the use of tools provided by computer-based technology. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society* (Proceedings of the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia, July 11-16). Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots8/ICOTS8_8D2_KAZAK.pdf

Konold, C. (2007). Designing a data analysis tool for learners. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 267-291). New York: Lawrence Erlbaum.

Konold, C., & Higgins, T.L. (2003). Reasoning about data. In J. Kilpatrick, W.G. Martin, & D. Schifter, (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 193-215). Reston, VA: National Council of Teachers of Mathematics.

Konold, C., & Miller, C.D. (2011). *TinkerPlots: Dynamic data exploration* [computer software, Version 2.2]. Emeryville, CA: Key Curriculum Press.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33***,** 259-289.

Lavigne, N.C., & Lajoie, S.P. (2007). Statistical reasoning of middle school children engaging in survey inquiry. *Contemporary Educational Psychology, 32,* 630-666.

Lehrer, R., Kim, M., & Jones, R.S. (2011). Developing conceptions of statistics by designing measures of distribution. *ZDM Mathematics Education*, *43*(5), 723-736.

Lehrer, R., Kim, M., & Schauble, L. (2007). Supporting the development of conceptions of statistics by engaging students in measuring and modeling variability. *International Journal of Computers for Mathematical Learning*, *12*, 195-216.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal, 41*(3), 635-680.

Makar, K. (2014). Young children's explorations of average through informal inferential reasoning. *Educational Studies in Mathematics*, *86*(1), 61-78.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning*, *13*, 152-173.

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

Meletiou-Mavrotheris, M., & Paparistodemou, E. (2015). Developing students' reasoning about samples and sampling in the context of informal inferences. *Educational Studies in Mathematics, 88*, 385-404.

Ministry of Education. (2007). *The New Zealand Curriculum*. Wellington, NZ: Author. Retrieved from http://nzcurriculum.tki.org.nz/The-New-Zealand-Curriculum

Mokros, J., & Russell, S.J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, *26*, 20-39.

Moore, D.S. (1990). Uncertainty. In L.S. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.

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

Moritz, J. (2000). Graphical representations of statistical association by upper primary students. In J. Bana & A. Chapman (Eds.), *Mathematics education beyond 2000* (Proceedings of the 23rd Annual Conference of the Mathematics Education Research Group of Australasia, Vol. 2, pp. 440-447). Fremantle, WA: MERGA. Retrieved from http://www.merga.net.au/documents/RP_Moritz_2000.pdf

Partnership for 21st Century Skills. (2009). *P21 framework definitions*. Retrieved from www.p21.org.

Peirce, C.S. (1931-1958). *Collected papers of Charles Sanders Peirce* (C. Hartshorne, P. Weiss, & A. Burks, Eds.) Cambridge, MA: Harvard University Press.

Rao, C.R. (1975). Teaching of statistics at the secondary level: An interdisciplinary approach. *International Journal of Mathematical Education in Science and Technology*, *6*, 151-162.

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *School and general issues* (Proceedings of the 3rd International Conference on the Teaching of Statistics). Voorburg, The Netherlands: International Statistical Institute. Retrieved from http://iase-web.org/documents/papers/icots3/BOOK1/A9-4.pdf

Shaughnessy, J.M. (2006). Research on students' understanding of some big concepts in statistics. In G. Burrill & P. Elliott (Eds.), *Thinking and reasoning with data and chance* (pp. 77-98). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J.M. (2007). Research on statistics learning and reasoning. In F.K. Lester, Jr. (Ed.), *Second handbook on research on mathematics teaching and learning* (pp. 957-1009). Charlotte, NC: Information Age Publishing.

Strauss, S., & Bichler, E. (1988). The development of children's concept of the arithmetic average. *Journal for Research in Mathematics Education, 19*, 64-80.

Thompson, H.A., Johnston, G., & Pfantz, T. (2006). Fish 'n' chips: A pedagogical path for using an in-class sampling experiment. In G. F. Burrill (Ed.), *Thinking and reasoning with data and chance* (pp. 449-465). Reston, VA: National Council of Teachers of Mathematics.

Van de Walle, J.A. (2004). *Elementary and middle school mathematics: Teaching developmentally* (5th ed.). Boston: Pearson.

Watson, J.M. (2007). The role of cognitive conflict in developing students' understanding of average. *Educational Studies in Mathematics*, *65*, 21-47.

Watson, J.M. (2009). The influence of variation and expectation on the developing awareness of distribution. *Statistics Education Research Journal*, *8*(1), 32-61.

Watson, J.M., Callingham, R.A., & Kelly, B.A. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning*, *9*, 83-130.

Watson, J., & English, L. (2015a). Expectation and variation with a virtual die. *Australian Mathematics Teacher, 71*(3), 3-9.

Watson, J., & English, L. (2015b). Introducing the practice of statistics: Are we environmentally friendly? *Mathematics Education Research Journal*, *27*, 585-613. DOI 10.1007/s13394-015-0153-z

Watson, J., & English, L. (2016a). *Do brown-eyed students have faster reaction times? The practice of statistics in Grade 6*. [Manuscript under preparation.]

Watson, J., & English, L. (2016b). Repeated random sampling in year 5. *Journal of Statistics Education*, *24*(1), 27-37. DOI: 10.1080/10691898.2016.1158026

Watson, J., & Fitzallen, N. (2016). Statistical software and mathematics education: Affordances for learning. In L. English & D. Kirshner (Eds.), *Handbook of International Research in Mathematics Education* (3rd ed.) (pp. 563-594). New York: Taylor and Francis.

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. (2000a). Developing concepts of sampling. *Journal for Research in Mathematics Education*, *31*, 44-70.

Watson, J.M., & Moritz, J.B. (2000b). The longitudinal development of understanding of average. *Mathematical Thinking and Learning*, *2*(1&2), 11-50.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248.

Zieffler, A., & Fry, E. (Eds.). (2015). *Reasoning about uncertainty: Learning and teaching informal inferential reasoning*. Minneapolis, MN: Catalyst Press.

JANE M. WATSON
Faculty of Education
University of Tasmania
Private Bag 66
Hobart, Tasmania 7001
Australia

Appendix A: Rules for “Quick Hands” Game

a) Each player’s hands must remain palms together

b) Player A’s finger tips must stay in contact with Player B’s finger tips

c) Player A may move/twitch his hands but they must remain in contact with Player B’s fingertips otherwise Player A forfeits his turn

d) Player A taps the back of the other player’s hand

e) Player B can withdraw her hands to avoid being tapped but only once Player A has broken the finger tip connection

f) A miss means it is the other player’s turn

Appendix B: Rules for Ruler Drop

a) Student A holds the ruler at the 30cm end, letting it hang down. He ensures the centimetre scale markings are facing Student B

b) Student B places her “dominant” hand at the bottom of the ruler (the 0cm end) without touching it

c) Student A announces he will drop the ruler in the next 5 seconds

d) When released, Student B catches the ruler between her thumb and index finger as quickly as possible

e) Read the measure on the ruler at the point the ruler is caught, that is measure under the thumb, rounding to the nearest whole centimetre

f) Record this measure in your Student Workbook below

g) The test may be repeated if needed due to a failure to conduct the test correctly (e.g. if a student completely misses/drops the ruler)

h) Swap roles and repeat the process with Student A catching

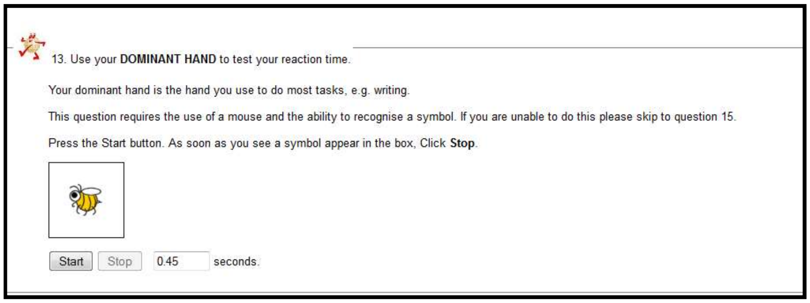
Appendix C: Instructions for ABS Reaction Timer

a) Use your dominant hand, just as you did for the Ruler Drop method

b) Each person gets one turn

c) Use the mouse pad to press “start” and then to press “stop” when the picture/symbol appears in the box

d) The timer will record a time in hundredths of a second, for example 0.45 seconds



13. Use your **DOMINANT HAND** to test your reaction time.

Your dominant hand is the hand you use to do most tasks, e.g. writing.

This question requires the use of a mouse and the ability to recognise a symbol. If you are unable to do this please skip to question 15.

Press the Start button. As soon as you see a symbol appear in the box, Click **Stop**.

0.45 seconds.

e) Record your reaction time below [in the Workbook]

Appendix D
Workbook Questions

Question Number	Question	Abbreviation
Q1/Q6	Part of analysing your data is making a representation of the data. *Represent* your data below. You may use any type of representation you like. Remember to ensure your representation helps you answer our question "What is the typical reaction time for Yr 5 students?"	**Ruler Plot/ ABS Timer**
Q2/Q7	Based on your analysis, What is the *typical reaction time* for Yr 5 students?	**Typical Time?**
Q3/Q8	*Explain* how you reached this conclusion. What tool/s did you use? What process did you follow?	**Explain**
Q4/Q9	How *certain* are you of your conclusion? Why?	**How Certain?**
Q5/Q10	Can you identify any *other issues* with the data collected or the tool/s or process used that might affect your conclusion?	**Other Issues?**
Q11A	Which of the two methods of determining reaction time is better/more reliable?	**Which Method Better**
Q11B	Why?	**Why Better?**

Appendix E: Rubrics for Coding Workbook Responses

Question	Code	Description	
Q1/Q6 Ruler Plot/ ABS Timer	0	No response	
	1	Incomplete plot or list, impossible to decipher	
	2	Unordered list or plot of data Data list not noting repeated values	
	3	Attempt to show calculation of statistic but incorrect	Value plot
	4	Ordered tallies of data, or frequency recorded	Horizontal or vertical stacked dot plot, or ordered bins
	5	Ordered tallies or list with indication of correct statistic (mean, median, mode)	Horizontal or vertical stacked dot plot with some marking to show a statistical value (mean, median, mode)
Q2/Q7 Typical Time?	**Need to code this in conjunction with the plot**		
	0	Number larger than 30 (length of ruler) / Whole number; decimal close to 0 or to 1.0 (ABS timer)	
	1	Number within the range of values in the plot	
	2	Number "near" middle of plot or repeated often (mode)	
Q3/Q8 Explain	0	No response, "guessed"	
	1	Non-specific response or description that does not fit mean, median or mode, generally "looked at the data"	
	2	Inconsistency of suggested method and suggested value (one reasonable and the other not)	
	3	Number seems roughly consistent with reasonable method (mean, mode, or median)	
Q4/Q9 How Certain?	0	No response, expression with no reason why	
	1	Certainty with vague, non-specific reason (data says so)	
	2	Certainty based on skill of calculating a value	
	3	Certainty based on meaning of tool (most, middle, "average")	
	4	Uncertainty because of small/class sample	
Q5/Q10 Other Issues?	0	No response, "None"	
	1	Data/computer wrong; errors in calculating; students "practice"	
	2	Wrong tools used	
	3	Sampling, only our class; repeated measures (for ABS timer)	
Q11A Which Method Better	**[Not hierarchical code]**		
	C	Computer	
	R	Ruler	
	S	Same	
	N	No response	
Q11B Why Better?	0	No response, unintelligible, no reason	
	1	Computer/ruler because more fun, more difficult to measure	
	2	Computer/ruler because more accurate, easier to measure	
	3	ABS, because it is time and Ruler is centimeters	