Mathematics Education Research Journal 2005, Vol. 17, No. 1, 24–57

Cognition and Instruction: Reasoning about Bias in Sampling

Jane Watson and Ben Kelly
Faculty of Education
University of Tasmania

Although sampling has been mentioned as part of the chance and data component of the mathematics curriculum since about 1990, little research attention has been aimed specifically at school students' understanding of this descriptive area. This study considers the initial understanding of bias in sampling by 639 students in grades 3, 5, 7, and 9. Three hundred and forty-one of these students then undertook a series of lessons on chance and data with an emphasis on chance, data handling, sampling, and variation. A post-test was administered to 285 of these students and two years later all available students from the original group (328) were again tested. This study considers the initial level of understanding of students, the nature of the lessons undertaken at each grade level, the post-instruction performance of those who undertook lessons, and the longitudinal performance after two years of all available students. Overall instruction was associated with improved performance, which was retained over two years but there was little difference between those who had or had not experienced instruction. Results for specific grades, some of which went against the overall trend are discussed, as well as educational implications for the teaching of sampling across the years of schooling based on the classroom observations and the changes observed.

Traditionally sampling was of minimal interest in statistics courses. Assumptions were made about random samples from normal distributions and then interest turned to hypothesis testing and confidence intervals. Sampling was more the domain of experimental scientists and social researchers. The aim was to be sure that the sample collected satisfied the criteria to hand the data over to the statisticians or computer programs to churn out statistics and p-values. The school curriculum reflected this approach in introducing students to the arithmetic mean in the middle years and the standard deviation, permutations, and combinations in the senior years. The advent of exploratory data analysis and its influence on the school curriculum since the National Council of Teachers of Mathematics' *Standards* (1989), have brought sampling to the forefront of the chance and data part of the mathematics curriculum. This perspective is reflected in *A National Statement on Mathematics for Australian Schools* (Australian Education Council [AEC], 1991) in band B for upper primary students with six types of activities to enable students to "understand what samples are, select appropriate samples from specified groups and draw informal inferences from data collected" (p. 172). Students are now expected to collect their own samples, explore the implications using descriptive statistics, and make judgements about claims, long before they are introduced to formal statistics

at the senior secondary level. It is hence important today for students to develop an appreciation of what sampling entails and to appreciate the similarities and distinctions between a statistical sample and a sample of food handed out in the supermarket.

In the classroom, however, anecdotal evidence suggests that sampling only gets passing mention. Unlike most parts of the mathematics curriculum, which require calculations to come up with specific answers, sampling is a topic described in words. Test questions would require answering in words not numbers and this sort of question is often unpopular with students and teachers alike. Sampling is more like a topic one would expect to find in a science course. It is to be hoped that some of the moves towards quantitative literacy (e.g., Madison & Steen, 2003) will help change these attitudes in the classroom and topics such as sampling will receive increased attention. In the transition time, research can address the issues of student understanding and ability to learn.

Previous Research

The early research with respect to understanding of sampling was related to the influence of sample size on decision-making. Tversky and Kahneman (1971) began with a study of college students and suggested that there was a tendency for them to believe that a sample, no matter how small, should represent the population exactly. They coined the term *representativeness heuristic* for this belief (Kahneman & Tversky, 1972) and spawned many studies related to judgments in situations of uncertainty. Although there has been some controversy about the complexity of the problems set by Tversky and Kahneman (Evans & Dusoir, 1977; Gigerenzer & Hoffrage, 1995) and in relation to whether questions are asked based on the centre of the distribution or the tail (Well, Pollatsek, & Boyce, 1990), it is generally acknowledged that issues of sample size and representativeness are important, particularly for students younger than those involved in these early studies.

Interest in school students' developing ideas of sampling has been considered from several angles, depending on the connections to other aspects of the chance and data curriculum that are considered important by researchers. Fischbein and Schnarch (1997), for example, used problems directly from Tversky and Kahneman (1971) with school students, whereas Estepa, Batanero, and Sanchez (1999) gave students specific pairs of data sets to compare as samples, and Reading and Shaughnessy (2000) asked students to imagine sampling in a probability setting. The specific issue of variability in sampling was considered for primary students by Wagner and Gal (1991) in the context of comparing two data sets of equal or unequal size; they found a dilemma for students between belief in homogeneity and anticipated variation. Rubin, Bruce, and Tenney (1991) found a similar tension for senior high school students in wanting both variation and representativeness in samples. Metz (1999), who worked with primary

students designing their own science experiments, observed a range of beliefs about sampling from "it really is not important" to "it is necessary to measure an entire population in order to reach decisions." In a more comprehensive analysis of these data (Metz, 2004), based on students' conceptualisations of uncertainty in scientific investigations, issues related to sampling were often discussed in relation to students' explanations for the uncertainty. The connection of representativeness to average was highlighted in the work of Mokros and Russell (1995), but not explicitly tied back to samples and sampling.

If representativeness is a quality to be sought in sampling, then bias is the other extreme and is to be avoided. Less research has focused specifically on this aspect. Jacobs (1997, 1999) worked with primary children and found that although in some situations they could identify potential sources of bias, they sometimes suggested spurious reasons for bias. They also experienced conflict in considering fairness, for example, in selecting a sample in relation to the desires of the people selected. Schwartz, Goldman, Vye, Barron, and The Cognition Technology Group at Vanderbilt (1998) observed similar results with school students of the same age. In a survey study of students' understanding of sampling, Watson and Moritz (2000a) found that 20% of grade 8 students could identify bias in at least one of two media contexts; by grade 11 this percentage rose to 66%. Based on in-depth interviews with 62 students in grades 3, 6, and 9, they found that one grade 6 and eleven (of 17) grade 9 students were sensitive to bias (Watson & Moritz, 2000b).

The basic definition of what constitutes a sample was considered by Watson and Moritz (2000a, 2000b) with the questions, "If you were given a 'sample,' what would you have?", and "Have you heard of the word 'sample' before? Where? What does it mean?" Four levels of response were observed reflecting the number of elements of relevance included in the description (from 0 to 3). The adequacy of this description was supported in later work of Watson and Kelly (2003) with a different group of students. Working with older students in an instructional setting, Saldanha and Thompson (2002) described different concepts of the sample-population relationship as being additive, where a sample is seen only in terms of the part-whole subset relationship, not as multiplicative, which also includes a "quasi-proportionality" relationship reflecting the features of the population. For students starting with more background in their study, these responses were parallel to the highest two levels observed by Watson and Moritz (2000a, 2000b).

Research Questions

As part of a larger study of school students' understanding of variability in relation to the chance and data curriculum and intervention to improve understanding, the following research questions were addressed in relation to the understanding of bias in sampling based on responses to survey questions.

1. What are the initial understandings of students in grades 3, 5, 7, and 9?
2. What change in understanding occurs after instruction in chance and data emphasising variation?
3. What level of understanding is sustained after two years for students who experienced instruction provided by the project and for those who did not? How do these compare?
4. How do students in longitudinal grades 5, 7, and 9 compare to the cohorts two years earlier?

Methodology

Sample

The sample presented here consists of 639 students from grades 3, 5, 7, and 9 in ten public schools in the Australian state of Tasmania who were surveyed as part of a larger study on school students' understanding of statistical variation including questions on basic chance, chance variation, data variation and sampling variation. Earlier analyses of these items have been reported in Watson and Kelly (2002a, 2002b, 2002c) for all grades. The sample sizes used in this study for each grade at each stage of the investigation are given in Table 1. The number of students in the current analysis is smaller than reported in the earlier analyses of Watson and Kelly, as the current analysis aims to deal with the questions related to sampling only.

Table 1
Number of Students in Each Grade

Grade	3/5[1]	5/7[1]	7/9[1]	9/11[1]	Total
Sample (Pre)	143	181	151	164	639
Sample (Post-Intervention)	57	80	76	72	285
Sample (Long.-Intervention)	36	53	51	23	163
Sample (Long.-Non Intervention)	47	35	53	30	165

[1]Grade in the longitudinal follow-up

Questions related to sampling were on the last half of the survey and some were not attempted by some students. To ensure a realistic data set on sampling, the authors deleted students who did not attempt at least two of the five "sets" of items in Figure 1 (Q1, Q2, Q3-Q8, Q9, or Q10-Q11) determined by physical placement on the survey.

Although not separated for the initial analysis in the study, 341 students were in schools where teaching intervention took place as part of the study

and 298 students were in schools with no intervention from the researchers. The reduction of the number of students from the pre-test to the post-test for students in the intervention schools was due to students having transferred from the school or being absent on the day the second survey was administered, approximately six weeks after the completion of the teaching intervention. The retention rate ranged from 76% to 91% over the four grades. For the longitudinal follow-up survey two years later, all grade 5 or 9 students in the same schools from grades 3 or 7 were surveyed again. For grade 7, all students from grade 5 in the associated feeder primary schools were surveyed. No transfers to other high schools were traced. Between grade 9 and 11, all students either left school or transferred to a regional senior secondary college. Students were traced to four regional senior schools for the final survey and the number was reduced mainly through students leaving school or not wishing to continue as part of the study.

The students surveyed were from ten schools considered to be typical of those in the state, each with a spread of academic ability. Five schools were in a relatively affluent area, three as feeder primary schools for two local high schools. The intervention high school had one intervention and one non-intervention feeder primary school. In the intervention high school, two grade 9 classes of "average" ability were assigned to the project, whereas the two grade 7 classes were of average to higher ability. In the non-intervention high school all students who were surveyed in grades 7 and 9 were of a range of ability levels. Five schools were in a less affluent area with three primary schools being feeder primary schools to two local high schools. The intervention high school had two intervention feeder primary schools. In the intervention high school at grade 9 three classes reflected different ability levels due to streaming of students, whereas the three grade 7 classes were of mixed ability. In the non-intervention high school there was a mix of ability levels in both grades 7 and 9. Five schools experienced instruction and five did not.

Tasks

The sampling questions shown in Figure 1 were part of a larger survey designed to assess school students' understanding of statistical variation in relation to the topics addressed in the chance and data curriculum. Questions Q1 to Q5, Q8, and Q9 were answered by students at all four grade levels, whereas Q6 was answered only by students in grades 5, 7, and 9; Q7, Q10, and Q11 were answered only by students in grades 7 and 9. Q6 and Q7 were omitted with younger students to shorten the time of administration of the survey, and Q10 and Q11 were only used with high school students because of the subject matter included. Q9 was the last part of a question related to reading information from a two-way table about participation of boys and girls in four sports at a school sports day. Q2 to Q8 were adapted from the work of Jacobs (1999), reflecting the standard accepted understanding of sampling appropriate for school-age students. Q10 and Q11 were used in an

earlier study by Watson and Moritz (2000a). These questions and the rest of the items in the complete survey administered in the larger study are discussed and analysed in terms of student understanding of statistical variation by Watson, Kelly, Callingham, and Shaughnessy (2003).

Q1. What does "sample" mean?
Give an example of a "sample".

Q2. A class wanted to raise money for their school trip to Movieworld on the Gold Coast. They could raise money by selling raffle tickets for a Playstation 2.

But before they decided to have a raffle they wanted to estimate how many students in their whole school would buy a ticket.

So they decided to do a survey to find out first. The school has 600 students in grades 1-6 with 100 students in each grade.

How many students would you survey and how would you choose them? Why?

Q3. **Shannon** got the names of all 600 children in the school and put them in a hat, and then pulled out 60 of them.

What do you think of Shannon's survey?

☐ GOOD ☐ BAD ☐ NOT SURE

Why?

Q4. **Jake** asked 10 children at an after-school meeting of the computer games club.

What do you think of Jake's survey?

☐ GOOD ☐ BAD ☐ NOT SURE

Why?

Q5. Adam asked all of the 100 children in Grade 1.

What do you think of Adam's survey?

☐ GOOD ☐ BAD ☐ NOT SURE

Why?

Q6. Raffi surveyed 60 of his friends.

What do you think of Raffi's survey?

☐ GOOD ☐ BAD ☐ NOT SURE

Why?

Figure 1 (cont.). Questions on sampling used in the survey.

Q7. Claire set up a booth outside of the tuck shop. Anyone who wanted to stop and fill out a survey could. She stopped collecting surveys when she got 60 kids to complete them.

What do you think of Claire's survey?

☐ GOOD ☐ BAD ☐ NOT SURE

Why?

Q8. Who do you think has the best survey method? Why?

Q9. A primary school had a sports day where every child could chose a sport to play. Here is what they chose.

	Netball	**Soccer**	**Tennis**	**Swimming**	**Total**
BOYS	0	20	20	10	50
GIRLS	40	10	15	10	75

a) How many girls chose Tennis?

b) What was the most popular sport for girls?

c) What was the most popular sport for boys?

d) How many children were at the sports day?

e) The teacher wanted to choose four children to lead the closing parade. Suggest two fair ways she could have chosen them.

The following article appeared in the Hobart *Mercury*.

Decriminalize drug use: poll

SOME 96 percent of callers to youth radio station Triple J have said marijuana use should be decriminalized in Australia. The phone-in listener poll, which closed yesterday, showed 9924 – out of the 10,000-plus callers – favoured decriminalisation, the station said.

Only 389 believed possession of the drug should remain a criminal offence. Many callers stressed they did not smoke marijuana but still believed in decriminalizing its use, a Triple J statement said.

Q10. What was the sample size in this article?

Q11. Is the sample reported here a reliable way of finding out public support for the decriminalisation of marijuana? Why or why not?

Figure 1. Questions on sampling used in the survey.

Procedure

The survey was administered in class time by the authors along with the classroom teachers, all offering help when required to read items, particularly in grades 3 and 5. Approximately 45 minutes was allocated for completing the surveys. The same survey was given to the same grade at each testing, hence for the two-year longitudinal survey, ex-grade 3 students

answered grade 5 questions and ex-grade 5 students answered grade 7 questions and so on. In comparing for longitudinal retention, only questions answered in the initial survey were counted, whereas all questions used in the final year could be used for cross-cohort comparisons.

Students in grades 3 and 5 in three of the six primary schools were taught a 10-lesson unit on chance and data emphasizing variation by a primary-trained mathematics specialist teacher provided by the project. The unit was taught over an 8-week period with two sessions at each school for each grade for each lesson. The content of the sessions is described in detail in Watson and Kelly (2002a) and summarized below. The students who received this intervention were administered the survey three times, initially (pre), six weeks after the instruction (post), and two years later (longitudinal). In the other three primary schools, there was no intervention from the project team. These students were only administered the survey twice, initially (pre) and two years later (longitudinal).

Session 1 of the 10-lesson unit used in the primary schools was an investigation of the contents of small packets of Smarties™. Beginning with a discussion of the information provided on the outside of the packet, students then worked in pairs to "find out" about the contents, creating column graphs of the Smarties™ sorted by colour. The discussion centred on the numbers of Smarties™ in each packet, the different colours, and the number of each colour in the individual packets. Variation among packets (as samples of the manufacturing process) was a focus of the class discussion as were the combined class data.

Session 2 aimed to develop ideas about defining the data to be collected, representing the data in different ways, and describing the general shape of the data. Data were collected, after suitable definitions were agreed to (differing from class to class), on the number of people in the children's families. Students themselves created people graphs and then used blocks before putting sticky dots on a class graph. There was discussion on the "most common" numbers in families, "outliers," and what could be said about half of the class, introducing the middle of the data. When describing the shape of the class graph, students tended to focus on individual features using terms like "chimney," but after encouragement to look more generally, began to use terms like "a mountain" or "a roller coaster" to describe variation observed.

The following two sessions were about chance, the first dealing with equally probable events using a spinner and a single die, and the second dealing with non equally probable events arising from the summing of outcomes when two dice are tossed. Students carried out repeated trials, recording outcomes and combining them as a class, again describing shapes of data, e.g., "a box" for a single die and "a hill" for summing two dice outcomes. The idea of gaining confidence when more data are collected was discussed. Sampling was the specific focus of the next two sessions with contributions of examples of samples from class members and overall

agreement on a basic definition. Selecting representative samples from the class population became a contentious issue when students agreed to random methods of selecting students, each of whom had an equal chance of being selected, but then were concerned about fairness if repeated sampling resulted in a student being selected a second time before everyone in the class had had a turn at being selected. Jacobs (1997, 1999) reported similar issues arising in her study. Sampling of cubes of two different colours from opaque bags, recording the results over many trials, and comparing the outcomes with expectation based on the bag's contents, were relatively sophisticated tasks for students in grades 3 and 5.

Sessions 7 and 8 were based on students' measurements of how long they could stand on one foot with their eyes closed (Rubin & Mokros, 1990). Again decisions were made on data collection and representation in order to compare two groups of data, for example, left and right feet, or boys and girls. The final two sessions allowed students to make decisions and set up their own investigations to answer questions related to blowing a pencil across a smooth surface. In all of the sessions collecting or describing data was a feature and variation in samples was pointed out at all points where it occurred. Although not always explicitly stated, "sampling" was a fundamental idea used throughout the unit of work with grade 3 and 5 students.

In grades 7 and 9, the regular mathematics teachers in two of the four secondary schools, delivered the unit of work to their classes, five classes in total in each of grades 7 and 9. The students in the other two secondary schools received no intervention from the research team. In the schools where there was instruction, there were nine different teachers involved, with one teacher taking two grade 7 classes. Because of the lack of control of exactly what would be taught and when, a comprehensive package of six small units of work, possibly encompassing several lessons, was prepared for the high school teachers. The topics included variation involved in trials of spinners, in outcomes with dice, in repeated sampling, in measuring association, in comparing two groups, and in the numbers of chocolate chips in cookies (Bright, Harvey, & Wheeler, 1981; Lappan, Fey, Fitzgerald, Friel, & Phillips, 1998; Lovitt & Lowe, 1993; Torok, 2000; Watson, 2002a). The sampling activities were similar to those used with grades 3 and 5, and the association and comparing groups units were based on the measurement of hand span and foot length. Although sampling was the specific focus of one unit, various types of samples were employed in all other units to collect data for analysis. Discussion of variation in sampling was hence intended to be widespread across the units.

The authors met with the grade 7 and 9 teachers in their schools, explaining the purpose of the project and distributing the 21 pages of lesson plans and 33 pages of associated documents (e.g., work sheets and copies of relevant pages from books). Suggestions as to the order in which the material might be taught were made, but final decisions were left to the judgments of the teachers. It was understood that the units were likely to be more

comprehensive than some teachers would be able to fit into their programs, although, as yet, no teachers had taught chance and data that year. As it turned out, there was considerable variation in the number of lessons taught by the high school teachers. Table 2 shows the number of lessons taught and the content of the lessons taught for each class in grades 7 and 9, with all teachers including at least one lesson on dice, but only one teacher touching on the chocolate chip cookie problem.

Table 2
Number and Content of Lessons Taught for Each Class in Grades 7 and 9

Grade/ Class	Unit 1: Spinners	Unit 2: Dice	Unit 3: Sampling	Unit 4: Association	Unit 5: Comparing Groups	Unit 6: Cookies	Total
7A	4	2	3	0	0	0	9
7B	4	2	3	0	0	0	9
7C	0	3	0	0	0	0	3
7D	1	3	0	1	0	0	5
7E	3	3	2	2	4	0	14
9F	0	3	1	2	0	0	6
9G	4	3	1	2	0	2	12
9H	1	1	0	1	0	0	3
9I	0	5	0	0	0	0	5
9J	3	4	4	0	3	0	14

As in the primary schools, students in the two secondary schools who received the intervention were administered the survey three times: initially (pre), six weeks after the instruction (post), and two years later (longitudinal). In the other two secondary schools, where there was no intervention, the students were administered the survey twice only: initially (pre), and two years later (longitudinal).

Analysis

Responses to the questions in Figure 1 were coded hierarchically to reflect an increasing appreciation and understanding of the concept of sample and bias in sampling. For Q1, which asked for a definition and an example of the term "sample", a Code 3 was given to responses that recognized the part-whole relationship along with the purpose to "test". Code 3 responses integrated all three of these appropriate ideas with examples (e.g., "A tester, an example, usually random, a small portion of the real thing. At a supermarket you might try a juice sample, a tiny cup so you can get a taste or idea"). A Code 2 response incorporated any two of the appropriate ideas in Code 3, be they the part-whole relationships (e.g., "A piece of something, water taken from the river is a water sample") or the purpose to test (e.g., "A little taste or try

of something, a small sample of chocolate"). A Code 1 was given to definitions with single ideas of either quantity (e.g., "A bit") or the purpose of a sample (e.g., "To try something"), or to examples only (e.g., "Blood sample"). A Code 0 was given to confused ideas of sample (e.g., "The trial run, the no's 1–6 on a dice"), or idiosyncratic responses, or no response to the question.

For Q2, which asked students to nominate how many people they would survey out of a school of 600 (100 in each of grades 1–6), and how they would choose them, five codes were developed from the four-code scheme used by Watson et al. (2003) to show an increasingly appropriate understanding of sampling methods. The highest code, a Code 4 was given to responses that combined an appropriate sample size with a random or representative and random method of selection (e.g., "I would randomly choose 10 students from every grade"); a Code 3 was given to the same kind of response, but with an unnecessarily large sample size (e.g., "I would choose 50 people out of each grade randomly picked"). Code 2 responses focused on representative methods of selection only, with either appropriate or inappropriate sample sizes (e.g., "I would ask 5 boys and 5 girls from each grade, which would make 60 students or 10%"), whereas Code 1 responses were non-representative and biased, regardless of the sample size given (e.g., "I would survey 10 students from each grade and pick them by whoever came first to volunteer"), or they focused on a method or on a sample size only, but not both in one response (e.g., "Ask people you see in the playground" or "100 of them"), or they wanted to survey the entire population (e.g., "I would survey the whole school, so that I would know exactly how many"). Code 0 was given to inappropriate responses, which misinterpreted the intent of the question, or to no response.

Q3 to Q7 were related to Q2, however, they provided different scenarios on how other hypothetical students decided to conduct their surveys of the school. The codes for each of the questions are presented in Table 3, adapted from Table 3 in Watson et al. (2003). Codes for Q3 to Q7 showed an increasing appreciation of sampling methods through critiquing others' methods. Code 3 responses provided an appropriate statistical critique of the method, Code 2 responses focused either on non-central issues regarding the method or on appropriate statistical issues with an element of uncertainty (noted by the "not sure" response). Code 1 responses focused on inappropriate issues, such as methods that create rather than remove potential bias, whereas Code 0 responses were idiosyncratic.

Table 3
Coding Categories of Response for Questions 3 to 7 (adapted from Watson et al., 2003)

Code	Q3 Shannon's method [all grades]	Q4 Jake's method [all grades]	Q5 Adam's method [all grades]	Q6 Raffi's method [grades 5-9]	Q7 Claire's method [grades 7-9]
3 – Appropriate statistical response	Random methods: "Good, because it's a good random way to survey"	Detecting bias and small sample size: "Bad, not enough people and selectively picked"	Detecting bias: "Bad, not enough different age groups"	Lack of range and/or variation: "Bad, they would probably say the same thing"	Appropriate criticism: "Bad, some kids might go twice"
2 – Non-central ideas or uncertainty	Adequate sample size: "Good, there's a lot of people"	Uncertainty: "Not sure, because not many different people would go there"	Large sample size: "Bad, too many people" Uncertainty: "Not sure, because that's only one class but he surveyed the most people"	Adequate sample size: "Good, you get a lot of answers" Uncertainty: "Not sure, it depends how many of his friends have different opinions"	Adequate sample size: "Good, you just have enough" Uncertainty: "Not sure, because people who thought it was a bad idea wouldn't bother"
1 – In-appropriate analysis	Method too random: "Bad, he could pick the wrong people"	Creating bias: "Good, to give them a hint to buy one"	Non-represent-ative: "Good, because it is fair"	Creating bias: "Good, because they are his friends"	Creating bias: "Good, it is their own choice"
0 – In-appropriate logic	Misinterpret question: "Bad, too many people"	Misinterpret question: "Good, so you could play it"	Misinterpret question: "Bad, none might not buy any"	Misinterpret question: "Good, more money for them"	Misinterpret question: "Good, first in best served"

Question 8, again related to Q2 to Q7, asked students to choose which survey method they thought was the best one and why. Although students in different grades were presented with a different number of potential survey methods (grade 3: Shannon's, Jake's, Adam's; grade 5: Shannon's, Jake's, Adam's, Raffi's; grade 7 and 9: Shannon's, Jake's, Adam's, Raffi's, Claire's), all grades were presented the statistically appropriate method to choose (Shannon's), and could therefore receive the highest code of 3 by choosing this method as the best one, combining it with a statistically appropriate reason (e.g., "Shannon, it was random and he doesn't ask his friends"). A Code 2 response also chose Shannon, but with an inappropriate reason or for no reason, or chose Shannon with another inappropriate choice (some responses said there were two best methods). A Code 1 was given to responses that focused on any of the other four options (or combination of options if two were selected) and provided an inappropriate statistical reason (e.g., "Claire, because it would get children from all ages and with different interests"), a reason based on fairness (e.g., "Claire, it's fairer"), or a methodological reason (e.g. "Adam, because he asked the most people and could times his results by six to get an average"). Code 0 was given to responses that were idiosyncratic or to no response.

Question 9 was a table-reading exercise about a sports day. Q9a) to Q9d) were basic table reading items and are not analysed here. Question 9e), however, focused on sampling and asked students to suggest two fair ways of picking children to lead a closing parade. For a code of 4, one out of the two responses focused on random and representative methods (e.g., "2 girls and 2 boys out of a hat"), or two responses were clearly distinct chance methods (e.g., "Pick out of a hat" and "Point to them without looking"); for a Code 3, at least one response had to be a simple chance method (e.g., "Put all the names in a hat and pull them out"). For Code 2 at least one response had to be a representative method by using two factors (gender/sport) (e.g., "Could have chosen 2 girls + 2 boys of which participated in everything"), and for a Code 1, at least one response needed to be representative using one factor (e.g., "Two boys and two girls"). Code 0 responses were again idiosyncratic or no response.

Codes for Q10, which involved reading the sample size straight from the article, were coded right-wrong, with Code 1 being for responses "10 313," "10 000+," or "10 000." Question 11, on the other hand required students to critique the sample method and claim of findings in the article. Four codes used by Watson and Moritz (2000b) were given to responses, with Code 3 responses giving multiple appropriate criticisms of the sampling method used (e.g., "No, not every one listens to Triple J and only the people who want to ring up will") and Code 2 responses focusing on single specific biases such as, the radio station (e.g., "No, only people [who listen to Triple J], because it's not random"), youth (e.g., "No, JJJ is a youth radio station, old people listen to Magic 107..."), and the response to phone polls (e.g., "No, because not all people will be bothered calling"). A code of 1 was given to a

variety of responses. Some code 1 responses critiqued the article for sample size issues, but without an appreciation of the part-whole relationships of sampling (e.g., "No, there are still heaps more people in Australia"), whereas others focused on the biases created by the type of callers phone polls attract (e.g., "No, because some could be users", "No, because some could lie"). Other Code 1 responses gave appraisal of the sampling process without recognition of the potential bias (e.g., "Yes, majority"). Code 0 responses were again idiosyncratic or misinterpretations, or were no response.

Except for Q10, which was coded 0/1, all questions had a maximum score of 3 (8 questions) or 4 (2 questions). The hierarchical rubrics produced an ordering for scores for questions from least to most statistically appropriately and when totalled the scores for grade 3 students could reach 23, for grade 5, 26, and for grades 7 and 9, 33. The scoring produced totals distributed as in most classroom testing, roughly normal with a slightly higher representation of 0 scores for grades 7 and 9 and a slight indication of skewness to the right for grade 3. Total scores of 0 represented responses to at least two of the five "sets" of items indicated in the section describing the sample.

T-tests were used to compare differences in means on total scores for the common items for each pair of grades (e.g., grade 3 and 5; grade 5 and 7; grade 7 and 9) for all students initially surveyed. Paired *t*-tests were used for comparing pre-test and post-test total scores for each grade, and to compare the pre-test and post-test total scores with the longitudinal follow-up for students who experienced intervention lessons. For the non-intervention students paired *t*-tests were used for pre-test and longitudinal total scores only. To consider potential differences in improvement for the intervention and non-intervention students, difference scores were calculated for each student and the means of these compared for the two groups with *t*-tests. *T*-tests were also used to compare mean total scores for each grade in the original year of testing with mean scores for the equivalent grade two years later in the longitudinal follow-up. Figure 2 highlights this last comparison. These last comparisons were carried out separately for intervention and non-intervention schools.

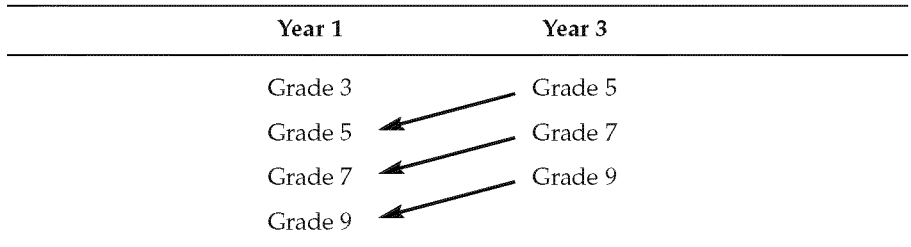

Figure 2. T-test comparisons for students in the final year of testing with students in the equivalent grades two years earlier.

Altogether 44 *t*-tests were performed and the conservative Bonferroni correction suggests a significance reduction from 0.05 to 0.0011. In the light of the information provided by the large number of *p*-values less than 0.05 (28) compared to the expected number (2.2), all *p*-values less than 0.05 are reported for consideration, however. The effect sizes for these differences were determined using Cohen's (1969) methodology and are reported with descriptors devised by Cohen (1969) and Izard (2004). The effect sizes were calculated using Coe's Effect Size Calculator (Lane, 2003).

Results

Descriptive Analysis of Items and Overall Performance by Grade in the Pre-test

For Q1 to Q7, Table 4 shows the overall percentage correct for each question. Question 1 was answered by all grades and asked for a definition of the term sample. The majority of students were able to give a single idea associated with the term or gave an example only (Code 1) but could not go further (40.2%), or were unable to define the term at all (32.5%). More grade 7 (12.6%) and 9 (11.6%) students reached the optimum level of response (Code 3), and roughly an equal percentage of students in grades 5 and 7 responded at Code 2 (18.2% and 21.2%, respectively). Grade 3 was the only grade to have a majority of Code 0 responses (63.6%), and the percentage of Code 0 responses declined over the grades to 29.3% in grade 5, 22.5% in grade 7, and 18.3% in grade 9. The modal response code for all grades except grade 3 was Code 1.

For Q2, more students responded at Code 1, citing non-representative methods and sample sizes, or they gave methods or sample sizes only, or wanted to survey the entire population. Very few students responded by giving random methods of selection (Codes 3 and 4) and no students in grade 3 gave appropriate sample sizes to match. The highest level of response for grade 3 was Code 3 (random methods without appropriate sample size). When answering this question, students seemed to find it difficult to formulate their own appropriate methods of sampling.

The next five questions, Q3 to Q7, asked students to critique proposed methods of sampling, most of them flawed with biases, with the exception of Q3. For Q3, which was the random method of Shannon, over half the students either did not respond, or did not give a reason for their decision. Of those who answered above Code 0, Code 1 was the modal response; however, there were almost as many Code 2 responses. Code 1 responses, in this case, were inappropriate criticisms to this method, focusing on the perceived inaccuracy of the random method, unfairness of opportunity, and small sample size. Code 2 responses reflected non-central appraisals, more specifically in relation to fairness and the sample size. Only 5.8% of students overall appropriately appraised this method citing "random" and/or

Table 4
Percentages of Responses for Each Code for Q1 to Q8 (all students who answered the question in the pre-test)

	Code 0	Code 1	Code 2	Code 3	Code 4
Q1	32.5	40.2	19.9	7.4	NA
Q2	27.5	51.2	14.5	4.8	1.9
Q3	54.6	20.5	19.1	5.8	NA
Q4	46.6	14.4	11.0	28.0	NA
Q5	39.3	20.5	23.5	16.7	NA
Q6[1]	39.9	18.5	33.3	8.3	NA
Q7[2]	44.1	43.2	7.9	4.8	NA
Q8	31.3	37.4	13.1	18.2	NA

[1] Answered by Grades 5 to 9 only
[2] Answered by Grades 7 and 9 only

"range" in their reasons, with an increase from 0% in grade 3 to 14% in grade 9. The percent of Code 0 responses dropped as grade increased, with a small rise from grade 5 to 7, dropping again at grade 9.

Students found Q4, Jake's method, easier to respond to than Q3; however, the modal response remained at Code 0. Students from each of grades 3 to 9 were able to detect the bias in the selection process, or mentioned the small sample size. The percentage of students responding at Code 3 rose monotonically from 7.7% in grade 3 to 45.7% in grade 9. There was a similar pattern of performance for Q5 on Adam's method; however, the modal response for those who answered above Code 0 was a Code 2. Code 2 responses focused on non-central criticisms, in this case the large sample size and fairness, or expressed some doubt in the criticism and were classified as statistically uncertain. The percentage of Code 3 responses per grade rose monotonically from 3.5% in grade 3 to 29.9% in grade 9.

Although only grades 5 to 9 answered Q6 about Raffi's method, the modal code of response was still Code 0 for this question. For the remainder of students who answered above Code 0, most responded at Code 2, again focusing on non-central criticisms (e.g., fairness) and non-central appraisals (e.g., sample size). Again, there was a monotonic rise in performance from grade 5 (2.2%) to grade 9 (14.6%) in Code 3 responses.

Only grades 7 and 9 answered Q7 based on Claire's method, with approximately 44% of students overall responding at Code 0. Broken down by grade, this accounted for 55% of grade 7 responses and 34.1% of grade 9 responses. The modal response above Code 0 was only Code 1 and most students gave inappropriate reasons for the method, based on perceived benefits of range and variation, fairness, and freedom of choice. Only 4.8% of grade 7 and 9 students were able to see the potential biases in this method of sampling.

Question 8 asked students to choose which they thought was the best method of sampling. Although all grades were presented with Q3, which was the statistically appropriate choice, the inclusion of Q6 and in particular Q7 for grades 7 and 9 had an impact on the results. Table 5 shows a decline by grade in the ability to choose the more appropriate method (Codes 2 and 3). A break down by grade reveals that 40.6% of grade 3 students and 46.4% of grade 5 students recognized Q3 as the appropriate method, whereas only 17.9% of grade 7 students and 21.9% of grade 9 students were able to do so. Although grade 5 students were not distracted by the inclusion of Raffi's method (Q6), the older students were distracted by Claire's method (Q7). Table 5 shows the percentages of students who chose each method in grades 3, 5, 7, and 9.

Table 5
Percentages of Responses for the Best Method for Q8 for all Students in the Pre-test

	Grade 3	Grade 5	Grade 7	Grade 9
Q3 (Shannon's method)	39.9	45.9	15.9	18.9
Q4 (Jake's method)	22.4	14.8	4.5	2.3
Q5 (Adam's method)	21.0	18.2	8.6	9.1
Q6 (Raffi's method)	NA	10.5	2.6	1.2
Q7 (Claire' method)	NA	NA	45.0	50.6
Q3 (Shannon) and other combination (2)	0.7	0.5	2.0[1]	3.0[1]
Q4, 5, 6, or 7 combination (2)		0.5		
All or none	1.4	2.2	4.6	3.7
Idiosyncratic, Don't know or no response	14.7	6.6	17.9	7.9

[1] Response denotes Shannon and Claire

Question 9, which asked for two fair ways to select students for a parade, provided a variety of responses. Overall, the modal response for this question was a Code 1 (36.9%), which required at least one method of selection that was representative based on one set of factors. The second most popular level of response, with 31.8%, was Code 3, which required at least one method of selection using simple chance methods, such as picking names out of a hat or rolling a die. By grade, the pattern of responses to Q9 was inconsistent. A Code 4 response, which required at least one random and representative method of selection, or two distinctly different chance methods of selection, was achieved by only 3.5% of grade 3 students, 3.9% of grade 5, 4.6% of grade 7, and 0.6% of grade 9 students. Similarly, 12.6%, 44.7%, 36.4%, and 29.9% of students in grades 3, 5, 7, and 9, respectively, achieved a Code 3 response. Code 0 responses made up 23.9% of the total responses, with grade 7 (32.4%) and grade 9 (36.6%) students contributing the most. This question was devised to tap into younger students'

knowledge of sampling; however, perhaps this disadvantaged older students who saw the question as too trivial to deserve a complex answer.

The results for Questions 10 and 11 were disappointing. Answered only by students in grades 7 and 9, 81.6% of students in Q10 were unable to identify the sample size from the article and were hence coded 0. Similarly, in Q11, no student in either grade managed to respond at Code 3, and only 11.7% responded at Code 2, and 15.2% at Code 1. The majority of responses (73%) were coded 0. Approximately 47% of students did not attempt both questions, as Q10 and Q11 were the last on the survey. This accounts in part for the large number of Code 0 responses, over half in Q10. Overall, grade 9 students performed slightly better than grade 7 students on both Q10 and Q11.

Performance on all questions for each grade is given in Table 6. As can be seen, the mean total scores are quite low in comparison with the maximum total score possible. There is, however, an increase in performance based on the percentages of the maximum total score from grade 3 to grade 5, with a drop from grade 5 to grade 7. There is another increase from grade 7 to grade 9; however, this percentage is no greater than the percentage of the maximum total score for grade 5 students.

Table 6
Mean Total Scores and Standard Errors for Each Grade on the Pre-test

		G3 (n=143)	G5 (n=181)	G7 (n=151)	G9 (n=164)
Pre-	Mean	5.52	9.57	9.93	12.08
	Std Error	0.285	0.336	0.514	0.468
	Maximum	23	26	33	33
	Mean as % of Maximum	24.0	36.8	30.1	36.6

Difference Between Grades

Mean total scores were also used to compare grades. Because grades 5 and 7 completed more questions than the next lower grade, totals were adjusted to include only those questions that were common for the lower grade. For example, for the grade 3/5 comparison, grade 5 responses for Q6 were not included in the total. Table 7 shows a significant difference between grades 3 and 5 and between grades 7 and 9 on the common items for the lower grade of each comparison, the greatest difference being between grades 3 and 5. For grades 5 and 7 there was a very small decrease in performance over the common items. There was a small difference between grades 5 and 9 (–1.81, p<.04) on these items.

Table 7
Mean Total Scores, Standard Errors, T-tests, and Effect Sizes for Adjacent Grades on the Pre-test

		G3 (*n*=143)	G5 (*n*=181)	G5 (*n*=181)	G7 (*n*=151)	G7 (*n*=151)	G9 (*n*=164)
Pre-	Mean	5.52	8.56	9.57	8.91	9.93	12.08
	Std Error	0.285	0.300	0.336	0.448	0.514	0.468
	t, *p*	–7.18, $p<.0001$		1.19, NS		–3.09, $p<.002$	
	Effect Size	0.80 (Large)		–0.13 (Very Small)		0.35 (Small)	

Pre-Post Analysis for Intervention Students

For the students who completed the post-test, paired *t*-tests were carried out for each grade. The mean total scores for the pre-test are reported again due to the reduced sample size. Table 8 shows that grades 3, 5, and 7 improved on the post-test after the teaching intervention to a small or medium extent; however, there was little improvement for grade 9 students. The similar means in grades 5 and 7 reflect the performance shown in Table 7 and the extra questions attempted in grade 7.

Table 8
Mean Total Scores, Standard Errors, Paired T-tests, and Effect Sizes for Each Grade on the Pre- and Post-test

		G3 (*n*=57)	G5 (*n*=80)	G7 (*n*=76)	G9 (*n*=72)
Pre-	Mean	6.60	9.48	9.49	12.06
	Std Error	0.497	0.479	0.629	0.638
Post-	Mean	8.00	12.13	13.01	12.61
	Std Error	0.670	0.570	0.808	0.795
	t, *p*	–2.38, $p<.02$	–5.01, $p<.0001$	–5.19, $p<.0001$	–0.70, NS
	Effect Size	0.31 (Small)	0.56 (Medium)	0.56 (Medium)	0.09 (Very Small)

In comparison with the pre-test percentages for each question that were given descriptively in the previous section and in Table 3, it is interesting to note that for the post- sample the pattern of post-test percentages showed an increase in the optimum level of response for every question. This increase in optimum responses was complemented by a decrease in the percentages of Code 0 responses to all questions. The percentages of students who selected the statistically appropriate method (Shannon) as the best method in Q8 revealed a mixed response in the post- sample, with a decrease for the grade 3 and 9 students and an increase for the grade 5 and 7 students. It is of

interest to note that this increase corresponds with the grades that showed the greatest improvement overall in the post-test after the instruction. There was, however, an increase in the percentages of students in grades 7 and 9 who inappropriately chose Claire as the best survey method in the post-sample.

Table 9 shows the pre- and post- means, standard errors, *t*-tests and numbers of lessons taught for the ten classes in grades 7 and 9. As can be seen, three of the grade 7 classes showed a significant increase in mean scores on sampling, whereas one class showed a slight improvement and another showed a significant decrease. It is interesting to note that the three classes that had a significant increase in mean score were also the classes who received lessons specifically in relation to sampling. The other two classes, although experiencing lessons that addressed variation through sample trials (e.g., dice, spinners), did not receive instruction specifically focused on sampling.

As seen in Table 9, three of the grade 9 classes also showed a significant increase in the mean score on sampling, whereas one class showed a slight improvement and another class had a significant decrease in performance. For the classes that showed a significant increase in understanding, two had experienced lessons specifically related to sampling (one each), with the third class experiencing none. Evidence from the teachers' journals suggests that the class that experienced no lessons was of a higher ability level than the other classes in that school; further, the class that received four lessons specifically in relation to sampling and showed a minor increase in understanding sampling was of a lower ability level. The class with the significant decrease in performance also did not receive any lessons focusing specifically on sampling.

Table 9
Pre- and Post- Means, Standard Errors, Paired T-tests and Total Number of Lessons for Each Class in Grades 7 and 9

Grade / Class	Pre Mean	Std Error	Post Mean	Std Error	*t*, *p*	Lessons
7A (n=17)	10.47	1.000	16.12	1.477	–4.17, p<.0004	9
7B (n=17)	12.00	1.331	17.53	1.292	–4.56, p<.0002	9
7C (n=9)	6.56	1.324	8.22	1.690	–1.69, NS	3
7D (n=9)	9.11	1.695	5.00	1.323	2.75, p<.02	5
7E (n=24)	8.25	1.296	12.42	1.371	–3.48, p<.002	14
9F (n=11)	13.27	1.251	16.45	1.592	–2.11, p<.04	6
9G (n=9)	7.67	1.826	12.78	1.211	–4.23, p<.002	12
9H (n=25)	13.80	1.112	16.36	1.139	–2.43, p<.02	3
9I (n=14)	12.79	1.314	4.86	1.440	5.17, p<.0001	5
9J (n=13)	9.92	1.337	10.38	1.328	–0.29, NS	14

Longitudinal Change

Table 10 shows the pre-test, post-test and longitudinal means, standard errors, *t*-test values and effect sizes for students who participated in the longitudinal follow-up in the schools with intervention. Again, the pre- and post- mean total scores are reported to reflect the reduced sample size. Even though each grade received the survey administered two years earlier to the equivalent grade (e.g., grade 3 students in 2000 now in grade 5 received the same survey as the grade 5 students in 2000), the mean total scores reflect what was achieved using only the items presented two years earlier.

Table 10
Mean Total Scores, Standard Errors, T-tests, and Effect Sizes for Each Grade on the Pre- and Post-test and the Longitudinal Follow-Up for the Students who Experienced Intervention

		G3/5[1] (n=36)	G5/7[1] (n=53)	G7/9[1] (n=51)	G9/11[1] (n=23)
Pre-	Mean	7.19	9.58	9.04	13.91
	Std Error	0.656	0.579	0.813	1.178
Post-	Mean	7.75	11.40	11.71	12.00
	Std Error	0.885	0.670	1.008	1.602
	t, *p*	–0.71, NS	–3.05, $p<.002$	–2.72, $p<.005$	1.17, NS
	Effect Size	0.12 (Very Small)	0.40 (Small)	0.41 (Small)	–0.28 (Small)
Long.	Mean	11.31	11.55	16.10	15.00
	Std Error	0.739	0.780	0.957	1.632
(pre-)	*t*, *p*	–5.20, $p<.0001$[2]	–2.73, $p<.005$[2]	–6.75, $p<.0001$[2]	–0.977, NS[2]
	Effect Size	0.98 (Large)	0.39 (Small)	1.11 (Large)	0.16 (Small)
(post-)	*t*, *p*	–4.86, $p<.0001$[3]	–0.20, NS[3]	–5.09, $p<.0001$[3]	–2.03, $p<.03$[3]
	Effect Size	0.73 (Medium)	0.03 (Very Small)	0.63 (Medium)	0.39 (Small)

[1] Grade in the longitudinal follow-up
[2] Pre-test to longitudinal follow-up
[3] Post-test to longitudinal follow-up

For each grade the effect size of change in the post-test decreased from that observed for the larger sample sizes reported in Table 8, with grade 9 showing a small decrease in performance. For this smaller group, after instruction there was an improvement (Pre- to Post-) for the grade 7 students, with a sustained (Pre- to Long.) and continued improvement over the two-year period (Post- to Long.). The grade 5 students also showed an improvement after the instruction and a sustained improvement long term over the two years (Pre- to Long.) but did not continue to improve after the

instruction like the grade 7 students (Post- to Long.). The specialized instruction had little effect on the grade 3 students (Pre- to Post-) but after two years the grade 3 students showed a large improvement. The grade 9 students who showed a small decrease after instruction (Pre- to Post-), reversed this to a small improvement after two years (Post- to Long.).

Examples to illustrate the increase in understanding after instruction (Pre- to Post-) with a sustained improvement over the two-year period (Long.) are given in Table 11 for the same individual students for each question. Although only 25% of students displayed this pattern of improved performance, it indicates what is potentially achievable.

Table 11
Examples of Improvement over the Three Survey Conditions for Selected Students

Question	Grade	Pre-test response	Post-test response	Longitudinal response
Q1 – Definition	7	"To test something out, some food or wine or something like that" (Code 1)	"A small amount of something. Something that has been tested" (Code 3)	"Sample means to take a bit of something and test it. Like trying a bit of bun at the bakery" (Code 3)
Q2 – Movieworld	7	"Make them all do the survey … 600 students … because they need money to buy tickets so the more people who know about it the better" (Code 1)	"Choose them randomly, 60 students in each grade, because [it] would give you enough people to get a good answer" (Code 3)	"I'd pick 20 people from each grade randomly, that should give a clear enough answer, [because] it makes sense" (Code 3)
Q3 – Shannon's method	9	"Good, because she knows how many would buy one" (Code 0)	"Good, because she picked people randomly" (Code 3)	"Good, because they were picked at random" (Code 3)
Q4 – Jake's method	5	"Good, because it will be good" (Code 0)	"Bad, he only had 10 people" (Code 3)	"Bad, only 10 people" (Code 3)

Q5 – Adam's method	5	"Good, he asked every single one" (Code 1)	"Bad, it is only getting one classes opinion" (Code 3)	"Bad, because he didn't get answers from all grades" (Code 3)
Q6 – Raffi's method	7	"Good, because if you X10 you get 600 and they would be the same age" (Code 1)	"Bad, because they would most likely [agree], they are his age" (Code 3)	"Bad, because they all might feel the same way" (Code 3)
Q7 – Claire's method	7	"Good, she's smart" (Code 0)	"Bad, because only people interested would do it" (Code 3)	"Bad, only people who would say yes would do it" (Code 3)
Q8 – Best method?	7	"Claire, because it is just a good idea" (Code 0)	"Shannon, because it's completely random" (Code 3)	"Shannon, because it was more random. She had a chance to get the whole school's opinion" (Code 3)
Q9 – Sports day parade	3	"2 girls, 2 boys" "Vote" (Code 1)	"Name out of a hat" "2 girls out of a hat, 2 boys out of a hat" (Code 4)	"Pulled name out of a hat" "Think of a number and the 4 people who guess closest go" (Code 4)
Q11 – Media	9	"Yes, because people could ring up and have a say" (Code 1)	"No, because it's not everyone, it's only the ones that listen to JJJ" (Code 2)	"No, generally only young people listen to JJJ so it isn't a fair sample group over the whole Australia" (Code 2)

Table 12 contains similar pre-test and longitudinal survey results to Table 10, for the non-intervention schools. Each grade showed some improvement in performance over the two-year period. The greatest improvement over two years was for grade 3. There was also a significant, yet smaller degree of improvement for students originally in grade 7 and grade 9. Although these students did not experience any intervention from the research project team, it is reasonable to expect some improvement over time due to the general school experience and maturation. For grades 3, 5, and 7, the effect size of the improvement from the pre-test to the longitudinal follow-up was not quite as great for the students in the non-intervention schools as it was for the students in the schools where the intervention took place.

Table 12
Mean Total Scores, Standard Errors, Paired T-tests, and Effect Sizes for Each Grade on the Pre-test and Longitudinal Follow-up for the Students who did not Experience Intervention

		G3/5[1] (n=47)	G5/7[1] (n=35)	G7/9[1] (n=53)	G9/11[1] (n=30)
Pre-	Mean	5.30	9.71	11.49	15.60
	Std Error	0.463	0.744	0.895	1.009
Long.	Mean	7.83	10.48	14.53	18.20
	Std Error	0.523	0.973	0.993	1.271
	t, p	–4.22, p<.0001	–0.93, NS	–3.71, p<.0003	–1.97, p<.03
	Effect Size	0.75 (Large)	0.15 (Small)	0.44 (Small)	0.41 (Small)

[1] Grade in the longitudinal follow-up

Comparison of Longitudinal Change for Intervention and Non-Intervention Schools

Table 13 contains the means, standard errors, two-tailed t-test results and effect sizes in comparing the difference scores (longitudinal – pre-test) for the intervention and non-intervention students at each grade level.

Table 13 shows that in grades 3 and 5 the intervention students had a higher mean difference score than the non-intervention students, but not significantly so. Although there is some indication that there was a greater positive difference for grade 7 students in schools with classroom intervention, the differences for other grades were negligible.

Change Within Schools Over a Two-Year Period

Detecting change within schools over the two-year longitudinal period was possible by comparing scores on common items for students originally in grade 5 in the first year of testing (pre-test in 2000), with students in grade 5

Table 13
Mean Total Scores, Standard Errors, Two-tailed T-tests, and Effect Sizes for Each Grade on the Difference Scores for the Students in the Intervention and Non-Intervention Schools

	Intervention (Mean, Std Error)	Non-Intervention (Mean, Std Error)	*t, p*	Effect Size
Grade 3/5[1]	4.11, 0.790	2.53, 0.600	1.62, NS	–0.36 (Small)
Grade 5/7[1]	1.96, 0.719	0.77, 0.826	1.07, NS	–0.23 (Small)
Grade 7/9[1]	7.06, 1.045	3.04, 0.817	3.04, $p<.002$	–0.60 (Medium)
Grade 9/11[1]	1.09, 1.112	2.60, 1.320	–0.84, NS	0.23 (Small)

[1] Grade in the longitudinal follow-up

(originally in grade 3) in the third year of testing (longitudinal follow-up in 2002) in both the schools that experienced intervention and the schools that did not. A similar comparison was carried out for students originally in grade 7 and grade 9 (see Figure 2 for clarification). Common questions were used for the comparisons.

For the schools that experienced intervention, Table 14 shows an improvement in performance for students in grade 5 in the longitudinal follow-up who had received instruction when they were in grade 3 two years earlier, when compared to the students originally in grade 5 before the intervention began. Similarly, students who were in grade 7 in the longitudinal follow-up who were originally in grade 5 and received instruction two years earlier, performed better than the original grade 7 students did. There was a non-significant improvement in favour of the longitudinal grade 9 students compared to those in grade 9 originally.

Table 14
Mean Total Scores, Standard Errors, T-tests, and Effect Sizes for the Same Grade Two Years Apart in the Intervention Schools

	Pre-test (2000) (Mean, Std Error)	Longitudinal (2002) (Mean, Std Error)	*t, p*	Effect Size
Grade 5	9.58, 0.579 (n=53)	12.75, 0.819 (n=36)	–3.25, $p<.0001$	0.70 (Medium)
Grade 7	9.04, 0.813 (n=51)	12.17, 0.827 (n=53)	–2.70, $p<.005$	0.53 (Medium)
Grade 9	13.91, 1.178 (n=23)	16.10, 0.957 (n=51)	–1.34, NS	0.34 (Small)

Table 15 shows that for the schools where the students did not experience intervention, there was no difference in performance between the students in grades 5, 7, and 9 in the longitudinal follow-up compared to the equivalent grades two years earlier. In fact in each case, there was a minimal drop in performance.

Table 15
Mean Total Scores, Standard Errors, T-tests, and Effect Sizes for the Same Grade Two Years Apart in the Non-Intervention Schools

	Pre-test (2000) (Mean, Std Error)	Longitudinal (2002) (Mean, Std Error)	*t, p*	Effect Size
Grade 5	9.71, 0.744 (n=35)	8.76, 0.599 (n=47)	1.00, NS	–0.22 (Small)
Grade 7	11.49, 0.895 (n=53)	11.26, 1.088 (n=35)	0.16, NS	–0.04 (Very Small)
Grade 9	15.60, 1.009 (n=30)	14.53, 0.993 (n=53)	0.70, NS	–0.16 (Very Small)

Discussion

The educational messages from this study are mixed. On one hand it is encouraging to observe significant change with a medium effect size in some instances after instruction, along with increases in the percentages of the highest level responses and decreases in the percentages of the lowest level responses to the items in the survey. On the other hand, the average performances across grades would not be considered satisfactory in terms of classroom learning objectives, as observed in the coding levels described for the items used in the surveys. Also, for students in grades 3 and 9, the effect size after the teaching intervention was small or very small, respectively. The outcomes are considered in more detail in relation to the research questions, the limitations of the study, and the educational implications.

Research Questions

The initial understanding of sampling showed a dip in performance by grade 7 students. This was evident both in the relative mean scores as a result of the maximum possible score on the questions asked (see Table 6) and in a comparison of grades on common items answered (see Table 7). In the latter case there was a small effect favouring grade 5 over grade 7, and the positive difference favouring grade 9 reflects to some extent the drop at the grade 7 level. There was only a small difference between grade 5 and grade 9, with a small effect favouring the grade 9 students. This dip in grade 7 performance was also observed in the larger study (Watson & Kelly, 2004) and has been seen in other studies of middle school students (Callingham & McIntosh, 2002; Hill, Rowe, Holmes-Smith, & Russell, 1996). As evidence continues to accumulate from studies across mathematics topics and other areas of the curriculum, the issue of the middle school drop in performance will require considerable attention.

The change in understanding observed after instruction was positive for each grade, although, as noted, the effect size was small for grade 3, and very small for grade 9. Students in grade 3 and grade 5 experienced the same lessons presented by the same teacher. Observation of the videotape of selected lessons indicated that specific discussion of sampling was a major

feature at both grade levels and it appeared that students were engaged in the tasks presented to them. It must be surmised that, in the short term, grade 3 students were unable to incorporate as much of the appreciation of sampling as the grade 5 students. As noted, there was less control by the researchers of the teaching that took place in grades 7 and 9. Anecdotal evidence (Watson & Kelly, 2002c) suggests that in each grade there was one classroom where the teacher experienced difficulty with the task, and that the grade 7 teacher who taught two classes was an enthusiastic participant in the project. It may also be relevant to recall that the grade 7 students started with a lower mean score than grade 5, and hence had quite a potential for improvement.

Over the two-year period of the project, each of the four grade levels in the intervention schools displayed a different pattern of improvement. The grade 5 students, who improved in the short term, were the only group not to show at least a small further positive effect after two years. This result is consistent with the middle school dip in performance observed in the initial data for grades 5 and 7. For grade 3 and grade 7 students the improvement over two years was impressive but little can be attributed to the short-term effect of the instruction for grade 3. For grade 7, the effect for the larger group of students who completed the post-test was more impressive than for the smaller group still in the study after two years, suggesting positive change in both the short and long term. For grade 9 students, small effects were seen in both the short and long term. Again, this improvement cannot be attributed with confidence to the short-term effect of the instruction experienced in grades 7 and 9 respectively. Furthermore, for both grade 7 and grade 9, there was a large degree of fluctuation for individual classes. The potential shown for improvement by the examples of individual students' responses to particular items is encouraging. The challenge is to help a larger group of students achieve such sustained improvement.

For the students in the schools that did not experience any intervention from the research team, the observed improvement over time was not surprising. "Chance and Data" has been a part of the Curriculum in Tasmania for a decade and it is assumed that the content is being taught in classrooms. It is known that several teachers in the non-intervention high schools attended Quality Teacher Programs, including sessions on chance and data led by the first author, at some stage after the initial testing; however, monitoring attendance at professional development seminars was not part of approved ethics procedures. The improvement for non-intervention schools highlights the need for caution when interpreting the long-term results of the students in the intervention schools, suggesting that the increased level of improvement in the longitudinal follow-up may have been due to other factors and not from the specific instruction implemented by the research team two years earlier.

Comparing grades 5, 7, and 9, at the end of the study, with their equivalent cohorts two years earlier in both the intervention schools and

non-intervention schools, suggested an encouraging result in that there was an indication that the teaching interventions for grades 3, 5, and 7, respectively, may have produced a better overall appreciation of sampling than was present at the same grade levels when the project began in the intervention schools. In contrast to this, there was no improvement in student understanding of sampling in the non-intervention schools in grades 5, 7, and 9 in the third year of testing, compared to the original students in grades 5, 7, and 9 in these schools.

Limitations

Several limitations of the project and its design should be acknowledged. This study itself was not based on a random sample. Such sampling is usually impossible in educational settings and certainly when a teaching intervention is involved. The schools chosen, however, were representative of the state government education system in Tasmania, and likely other state systems in Australia.

The control over the teaching sequence in grades 7 and 9 was much more limited than in grades 3 and 5. Providing a high school mathematics teacher for all classes within a complex high school timetable was beyond the financial resources of the project. It was also expected by the researchers that although primary school teachers might be intimidated by elements of the chance and data curriculum, high school mathematics teachers should not be. This may have been a misapprehension, particularly in terms of teachers' motivation to teach and enthusiasm about the topics. As reported in Watson and Kelly (2002c) for the ten classes in grades 7 and 9, the overall correlation of number of lessons taught and the "post- – pre-" mean score on the larger survey of which the sampling subscale used in this study was a part, explained only 18% of the variance and was not statistically significant. Hence the number of lessons taught cannot be hypothesized as a predictor of motivation on the part of teachers to enthuse the students or of students' greater achievement, either overall or in relation to sampling. Helme and Stacey (2000) encountered a similar situation when they provided resources for teaching decimals to four willing primary teachers, with only one consistently using them. In their study, student outcomes were strongly related to teacher use of materials, a result not as evident in the current study that involved high school teachers. As noted earlier, in one high school, the grade 9 classes selected were said by the organising person to have students of "average" ability, rather than a wide range of students including higher ability. Although this was catered for in terms of pre-test and post-test measurements, again it may have influenced the interest and motivation of the students.

Although sampling was the focus of some lessons, and discussion about samples took place in all grade 3 and 5 classes and it is assumed in most grade 7 and 9 classrooms, except for the definition of sample itself, there was no specific reference during teaching to the items on the survey. In particular,

the Movieworld questions (Q2 to Q8 in Figure 1) were intended to be of a sufficiently general and familiar nature that they would measure the transfer of understanding from activities carried out in the classroom. It may be that the discussion of bias in sampling in the classroom was not sufficiently similar to the context of the items in the survey to encourage transfer. Pertinent to this question is the item about Claire's method (Q7 in Figure 1), which was only given to students in grades 7 and 9. As seen in Table 5, the presence of this item was a major distracter for these grades in determining which sampling method was the most appropriate.

As noted in the description of the coding system leading to the scores used for measuring student performance, the rubrics represented the authors' views of statistical appropriateness for students at the school level. The hierarchical nature of the scoring is inherent in this appropriateness and in the increasingly complex structure observed in the responses (Pegg, 2002). Others may have a different perspective on coding.

From a measurement perspective the presence of Claire's method (Q7) disadvantaged the students in grades 7 and 9 compared to grades 3 and 5, and may contribute marginally to their smaller increase in performance levels with respect to the earlier grades. The presence of the question from an educational standpoint, however, provides valuable information about students' beliefs concerning sampling. Ideas of fairness in a colloquial sense, and allowing for voluntary participation, are more important to students than avoiding bias by using a random method. Teachers need to be aware of these beliefs and make specific provision for discussing them in the classroom.

Educational Implications

Although this intervention study sought to compensate for the researchers' perception that previous teaching in relation to the chance and data curriculum had neglected specific descriptive discussion of sampling, it is clear that even more needs to be done along these lines. Even though teachers may emphasize the importance of understanding samples and the purpose of avoiding bias, students may not appreciate the importance and lose concentration because numbers and calculations are not being presented to them, as is the perception of a normal mathematics classroom. The discussion of Jacobs (1999) is helpful in this regard for thinking of students in the upper primary years. As well as stressing the need to confront students by challenging the "fairness" and "self-selection" rationales, she suggests two further considerations for designing instructional activities. First, she suggests that teachers need to give students practice at making decisions from the results of multiple surveys, as students tend to aggregate information from all surveys when drawing conclusions, even after identifying biases with certain methods. Second, teachers should supply students with surveys based on multiple situations in a variety of contexts, including within the school and in the outside world. Surveys conducted

outside the school context often result in students seeing more clearly the reason to use samples due to the larger population and the inability to survey everyone.

Carrying out sampling activities in the classroom as suggested by Watson and Shaughnessy (2004) in the context of drawing handfuls of lollies from a container with a given percentage of a certain colour, can also be useful. Students' discussion of their own methods of drawing handfuls is likely to bring out accusations of cheating or bias on the part of other students. Activities such as this one link to the chance part of the curriculum in terms of predicting outcomes based on the proportion of each colour present in the container. Repeated sampling (with replacement) from mystery bags containing a small number of coloured objects, with the aim of guessing the number of objects of each colour, is another activity (used in some classes in this study) that can reinforce appropriate ideas of sampling technique and sample size. Watson (2002b) describes the bias that occurs when data from two samples of size two are combined as if they were one sample of size four. Allowing students to experience such difficulties and discover the consequences may be instrumental in building appropriate understandings of the sampling process. It is also possible to introduce students to the interesting history of the development of sampling methodology within the field of statistics (e.g., Bernstein, 1998; Salsburg, 2001). To hear of the difficulties and debates experienced over the past two centuries may help students to appreciate their own dilemmas in considering bias in sampling. It will also help them prepare for more advanced work where subtle issues of sampling are considered in more detail than is possible in the middle and high school years.

The poor performance of students in grades 7 and 9 on questions related to an article about a survey from the media was due, in part, to students not being able to finish the survey in the timeframe that was allowed; however, since 25% of students responded in an idiosyncratic manner, it may also be a reflection on student unwillingness to "read" questions on what is perceived to be a mathematics test. Further, it may be related to low literacy levels or to a lack of experience with critical reading of the newspaper. As noted elsewhere (Gal, 2002; Watson, 1997, 2000) the ability to read and question media articles is an important constituent of the statistical literacy needed by students when they leave school. Learning to question sampling procedures as presented by the media is an important part of this ability. Its importance is recognised in the Australian *National Statement* (AEC, 1991) in a specific activity for students, "Discuss and make judgments about arguments and claims in the media for which statistical information is presented (e.g. claiming that 40% of the community think that the school leaving age should be raised on the basis of a telephone 'ring-in' poll)" (p. 172).

The results of this study suggest that more research is needed into intervention programs that seek to improve students' understanding of sampling and associated bias within the chance and data curriculum. The

use of student interviews, both initially (e.g., Watson & Moritz, 2000b), and longitudinally (e.g., Watson, 2004), to supplement information from surveys, is likely to assist in the further development of materials and teaching techniques to improve understanding. Carrying out interviews during the teaching intervention itself, is another potential aid, along with greater liaison with teachers during this time. Results of this study suggest that the focused interaction of researchers, teachers, and students during a planned intervention is likely to produce the greatest benefit in relation to long-term outcomes.

Acknowledgements

This research was funded by Australian Research Council grants (No. A00000716 and No. DP0208607). The authors wish to thank Patricia Jeffery, the project's classroom teacher for grades 3 and 5.

References

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

Bernstein, P. L. (1998). *Against the gods: The remarkable story of risk*. New York: Wiley.

Bright, G. W., Harvey, J. G., & Wheeler, M. M. (1981). Fair games, unfair games. In A. P Shulte & J. R. Smart (Eds.), *Teaching statistics and probability: 1981 Yearbook* (pp. 49–59). Reston, VA: National Council of Teachers of Mathematics.

Callingham, R., & McIntosh, A. (2002). Mental computation competence across years 3 to 10. In B. Barton, K. C. Irwin, M. Pfannkuch & M. O. J. Thomas (Eds.), *Mathematics education in the South Pacific* (Proceedings of the 25th annual conference of the Mathematics Education Research Group of Australasia, Auckland, Vol 1, pp. 155–163). Sydney: MERGA.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.

Estepa, A., Batanero, C., & Sanchez, F. T. (1999). Students' intuitive strategies in judging association when comparing two samples. *Hiroshima Journal of Mathematics Education, 7*, 17–30.

Evans, J. St. B. T., & Dusoir, A. E. (1977). Proportionality and sample size as factors in intuitive statistical judgement. *Acta Psychologica, 41*, 129–137.

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28*, 96–105.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*, 1–51.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.

Helme, S., & Stacey, K. (2000). Improving decimal understanding: Can targeted resources make a difference? 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. 1, pp. 299–306). Perth, WA: MERGA.

Hill, P. W., Rowe, K. J., Holmes-Smith, P., & Russell, V. J. (1996). *The Victorian Quality Schools Project: A study of school and teacher effectiveness. Report (Volume 1).* Melbourne: Centre for Applied Educational Research, University of Melbourne.

Izard, J. F. (2004, March). *Best practice in assessment for learning.* Paper presented at the Third Conference of the Association of Commonwealth Examinations and Accreditation Bodies on *Redefining the Roles of Educational Assessment*, South Pacific Board for Educational Assessment, Nadi, Fiji.

Jacobs, V. R. (1997, March). *Children's understanding of sampling in surveys.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgement of representativeness. *Cognitive Psychology, 3*, 430–454.

Lane, D. M. (2003). Hyperstat on-line textbook. Retrieved October 21, 2004, from http://davidmlane.com/hyperstat/index.html

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., & Phillips, E. D. (1998). *Samples and populations.* Menlo Park, CA: Dale Seymour Publications.

Lovitt, C., & Lowe, I. (1993). *Chance and data investigations* (Vols. 1–2). Melbourne: Curriculum Corporation.

Madison, B. L., & Steen, L. A. (Eds.). (2003). *Quantitative literacy: Why numeracy matters for schools and colleges.* Princeton, NJ: The National Council on Education and the Disciplines.

Metz, K. E. (1999). Why sampling works or why it can't: Ideas of young children engaged in research of their own design. In F. Hitt & M. Santos (Eds.), *Proceedings of the 21st Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 492–499). Cuernavaca, Mexico: PMENA.

Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualisation of uncertainty in investigations of their own design. *Cognition and Instruction, 22*(2), 219–290.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20–39.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Pegg, J. E. (2002). Assessment in mathematics: A developmental approach. In J. M. Royer (Ed.), *Mathematical cognition* (pp. 227–259). Greenwich, CT: Information Age Publishing.

Reading, C., & Shaughnessy, M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89–96). Hiroshima, Japan: Hiroshima University.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. 1, pp. 314–319). Voorburg, The Netherlands: International Statistical Institute.

Rubin, A., & Mokros, J. (1990). *Data: Kids, cats, and ads: Statistics.* Menlo Park, CA: Dale Seymour Publications.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257–270.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: Henry Holt.

Schwartz, D. L., Goldman, S. R., Vye, N. J., Barron, B. J., & The Cognition and Technology Group at Vanderbilt. (1998). Aligning everyday and mathematical reasoning: The case of sampling assumptions. In S. P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching and assessment in grades K–12* (pp. 233–273). Mahwah, NJ: Lawrence Erlbaum.

Torok, R. (2000). Putting the variation into chance and data. *Australian Mathematics Teacher, 56*(2), 25–31.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105–110.

Wagner, D. A., & Gal, I. (1991). *Project STARC: Acquisition of statistical reasoning in children*. (Annual Report Year 1). Philadelphia, PA: University of Pennsylvania, Literacy Research Center.

Watson, J. M. (1997). Assessing statistical literacy using the media. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107–121). Amsterdam: IOS Press and The International Statistical Institute.

Watson, J. M. (2000). Statistics in context. *Mathematics Teacher, 93*, 54–58.

Watson, J. M. (2002a). Lessons from variation research II: For the classroom. In M. Goos & T. Spencer (Eds.), *Mathematics – making waves* (Proceedings of the Nineteenth Biennial Conference of the Australian Association of Mathematics Teachers Inc., Brisbane, pp. 424–432). Adelaide: AAMT Inc.

Watson, J. M. (2002b). When 2 + 2 □ 4 and 6 + 6 □ 12 in data and chance. *New England Mathematics Journal, 34*(2), 56–68.

Watson, J. M. (2004). Developing reasoning about samples. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 277–294). Dordrecht, The Netherlands: Kluwer.

Watson, J. M., & Kelly, B. A. (2002a). Can grade 3 students learn about variation? In B. Phillips (Ed.), *CD of the Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society*, Cape Town, South Africa. Voorburg, The Netherlands: International Statistical Institute.

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

Watson, J. M., & Kelly, B. A. (2002c). Variation as part of chance and data in grades 7 and 9. In B. Barton, K. C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds.), *Mathematics education in the South Pacific* (Proceedings of the 26th annual conference of the Mathematics Education Research Group of Australasia, Auckland, Vol. 2, pp. 682–689). Sydney: MERGA.

Watson, J. M., & Kelly, B. A. (2003, December). *The vocabulary of statistical literacy*. Refereed paper presented at the joint conferences of the New Zealand Association for Research in Education and the Australian Association for Research in Education , Auckland, New Zealand. Retrieved March 20, 2005 from http://www.aare.edu.au/03pap/wat03297.pdf

Watson, J. M., & Kelly, B. A. (2004). A two-year study of students' appreciation of variation in the chance and data curriculum. In I. Putt, R. Faragher, & M. McLean (Eds.), *Mathematics education for the third millennium: Towards 2010* (Proceedings of the 27th Annual Conference of the Mathematics Education Research Group of Australasia, Townsville, Vol. 2, pp. 581–588). Sydney: MERGA.

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

Watson, J. M., & Moritz, J. B. (2000a). Development of understanding of sampling for statistical literacy. *Mathematical Behavior, 19*, 109–136.

Watson, J. M., & Moritz, J. B. (2000b). Developing concepts of sampling. *Journal of Research in Mathematics Education, 31*, 44–70.

Watson, J. M., & Shaughnessy, J. M. (2004). Proportional reasoning: Lessons from research in data and chance. *Mathematics Teaching in the Middle School, 10*, 104–109.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes, 47*, 289–312.

Authors

Jane Watson, University of Tasmania, Private Bag 66, Hobart, Tsmania, 7001.
Email: <Jane.Watson@utas.edu.au>

Ben Kelly, University of Tasmania, Private Bag 66, Hobart, Tsmania, 7001.
Email: <Ben.Kelly@utas.edu.au>