Randomness, sample size, imagination and metacognition

making judgments about differences in data sets

Sue Stack
University of Tasmania
<susan.stack@utas.edu.au>
Jane Watson
University of Tasmania
<jane.watson@utas.edu.au>

Consider this:

> An experimental treatment for mildly to moderately depressed patients is tested by taking two groups of 15 to the Caribbean. One group swims 4 hours per day for 4 weeks (the control group) and the other group swims for 4 hours per day for 4 weeks with dolphins (the treatment). Ten out of 15 of the dolphin group improve, whereas 3 out of 15 of the swimming only group improve (Rossman, 2008).
> What could you claim from this experiment and for what population? How certain would you be? Why? What types of things are you thinking about to inform your decision? What would you like to know more about? What are the big statistical ideas in this problem?

Informal inference
- based on evidence
- can be generalised to a larger population
- gives an estimate of certainty.

This was one of the problems that Grade 10 mathematics students were challenged to consider as a part of an action research project, where two teachers trialled a statistics and probability unit specially developed by the second author. A key intent of the five week unit was to help students develop conceptual and technical tools to be able to make informal inferences (statistical judgments) about whether claims based on one sample could be generalised to larger populations, and about whether differences in data sets were significant or whether the differences might have occurred by chance (a random event). Through using the TinkerPlots program (Konold & Miller, 2011) students could collect samples and create simulations that would enable them to develop visual and intuitive understandings alongside statistical interpretations.

Random phenomenon
A phenomenon is called random if the outcome of a single repetition is uncertain but there is nonetheless a regular distribution of relative frequencies in a large number of repetitions.

Simple random sample
A simple random sample of size *n* consists of *n* units from the population chosen in such a way that every set of *n* units has an equal chance to be the sample actually selected.

(Definitions from Moore and McCabe, Introduction to the Practice of Statistics,1993)

The lead teacher was keen to expose her students to technical tools that might expand their notion of what mathematical inquiry is and liked the way that the unit combined ideas of probability and statistics, concerned that the normal way of teaching probability was around gaming contexts.

A key challenge of the project was building an understanding of the concept of randomness within the activities in such a way that students could draw meaningful conclusions about populations from small samples. The conversations that the two teachers and the two researchers had with students problematised the existing understandings of what randomness was—it highlighted the range of different meanings and contexts, and how and when they are used. As a result they became active inquirers into concepts around randomness, realising that in building student capacity for inference it was crucial explicitly to develop and distinguish different ideas around randomness.

Based on the activities used with the students, one aim of this article is to tease out some of the aspects of randomness that emerged as a result of this project that might be useful for other teachers. A metacognitive frame for thinking about informal inference is introduced, which takes into account the thinking that students were actually engaged in during the unit of work. It was realised how important the role of imaginative thinking is in helping students to engage with the context, imagine alternative hypotheses, imagine the probability space, imagine their data in new ways and imagine whole populations.

Swimming with dolphins

Take a moment to imagine the issues in the dolphin problem. Does it help to imagine being there in the Caribbean swimming with dolphins? What might be the *reasons* that swimming with dolphins is a positive experience (special experience, bonding, dolphins are friendly) and why might it be negative (scary, not in control)? Does it help to imagine how the scientists may have *designed the experiment*? Did they select the 30 people randomly from an entire population of depressed people, or did they invite people to participate? Would a randomly selected sample be representative? How would we know it was representative? Would inviting people to participate tend to select people who like dolphins and therefore the experiment would be difficult to generalise to a larger population? How was the control group of 15 people selected out of the 30 people? Was it done randomly and why might this give a fairer test? What constitutes an improvement? How long was the improvement sustained after the tests?

Chance devices	Measurement contexts
Single random outcome Expect an unpredictable outcome: • Toss of a coin. Could be either head or tail. • Roll of dice. Could be 1, 2, 3, 4, 5, or 6.	**Random sample size of 1 from a population** Expect an unpredictable outcome. Unknown likelihood to be representative of the population.
Small random sample of outcomes Expect variation in outcomes from the probability model. Repeating these samples will show high variation between the samples. Toss a coin 10 times. Might have 7 heads and 3 tails. Repeating this five times might have the following results: 6/10H, 4/10H, 7/10H, 5/10H, 4/10H.	**Small random sample** Expect that small samples are likely to vary from the population. Comparing different small samples of the same sample size, each randomly selected from a population, will show variation among them.
Large random sample of outcomes Expect a pattern to appear reflecting the underlying probability model. • 1000 coin tosses. Expect percentage of heads to be close to 50% (theoretical probability). • 1000 dice throws. Expect relative frequency of each number to be close to 1/6 (theoretical probability).	**Large sample size randomly selected** Expect close agreement with the population. Comparing different large random samples will show little variation among them.

Now look at the data: Sixty-seven per cent of people improved when swimming with dolphins and 20% of people improved just swimming. This seems to indicate that swimming with dolphins does make a difference. But could this result have happened by chance? Perhaps it was an extreme random event. The sample is very small—only 15 people in each group. In small samples would we expect more extreme events? How could we test whether this is the case for these data?

How many different ways have we considered randomness in this problem? What student experience and understanding of randomness are we assuming?

It is useful to consider randomness in two key mathematical contexts: randomness as used in *chance devices* (where it is theoretically possible to generate infinite populations from single trials), and randomness in *measurement contexts* (where there is a finite population of fixed yet undetermined values that sampling aims to discern). How does randomness behave in these different contexts and can understanding randomness in one context help to understand it in the other?

Randomness and the long term

In the action research project students initially used 'random' as a slang word meaning 'unexpected'. 'It is random' meant it could be anything.

A key idea for chance devices is that in the long term (for large samples) we will see an emerging pattern of outcomes that reflects the underlying theoretical probability model. This is a model that can be deduced through imagining all of the possibilities in the probability space. Research suggests that as a result of playing games few students are likely to believe that in the long term random generation ensures fairness (Watson & Moritz, 2003). In the short term (or for small samples) there is an unpredictability and it is possible that extreme values occur. The result may not appear fair or representative at all. Compare 30 rolls of a die versus 3000 based on random generation of data in *TinkerPlots* in Figure 1.

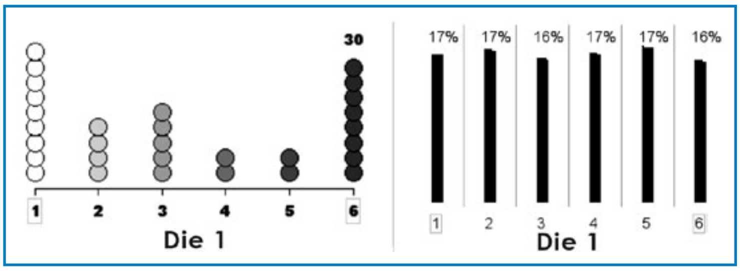


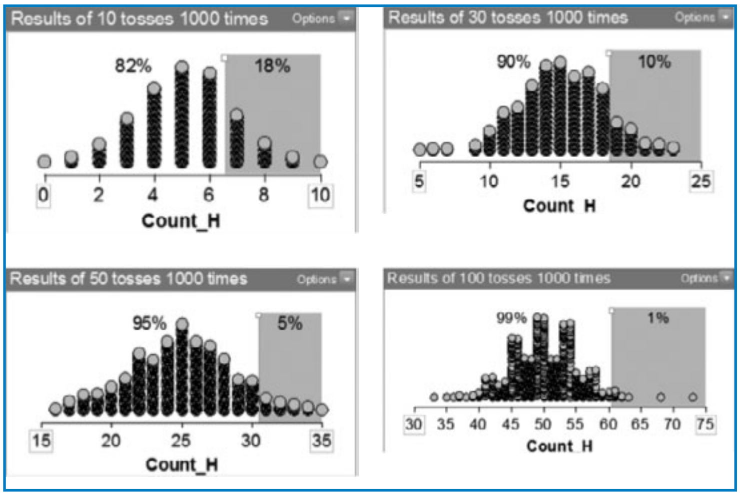
Figure 1. Outcomes from rolling a die 30 times or 3000 times.

The unusual outcome on the left of Figure 1 might lead us to believe it was caused as a result of a specific intervention (i.e., using a loaded die), when in fact it might just be a "usual" characteristic of randomness—the generation of an extreme event. Research suggests that students have difficulty in discerning between caused and random extreme events generated by chance devices (Chiesi & Primi, 2009). Young students for example, seeing 5 heads in a row might be likely to predict another Head, thinking it is a caused pattern, whereas older students might predict a Tail because they expect the results to even out, not recognising that each toss is an independent event. A typical assumption by students is that a large sample of random events will be replicated in a small sample (Watson & Moritz, 2000). It is difficult for many to imagine how the sample size might affect the number of extreme events that we see.

Students expected to see the frequency distribution of the die to even out in the long term, but were surprised by how many rolls of the die it took.

Comparison of sample sizes in chance devices

Using simulation we can compare a range of sample sizes to generate random data. We toss a coin 10 times to represent a sample size of 10, and compare that with 30, 50 and 100 tosses, to represent larger sample sizes (see Figure 2). For



Some of the students predicted the larger sample would have the more extreme values, a few thought no difference, and a few nominated the small sample. Those who initially predicted that more extreme values would be found in the larger sample were convinced after generating their own sample comparisons that the smaller sample created more extreme events.

Figure 2. Number of heads in 1000 trials of 10, 30, 50 or 100 tosses of a coin (shaded area shows more than 60% heads).

each one we do this for 1000 trials to get a sense of what the pattern will look like in the long term. The plots show the frequency of outcomes for heads. The students were interested in seeing how often heads occurred more than 60% of the time in each set of samples—these outcomes represented an ‘extreme’ event. As the sample size got larger the percentage of extreme events reduced.

Comparison of sample size for fixed populations

The previous simulation is a very visual demonstration about how sample size in the context of chance devices affects the generation of extreme events. We would expect a similar trend comparing a range of sample sizes of large finite populations—the smaller the sample size the greater the chance of extreme events.

In the following situation we have a known population of 677 data points. Each data point represents a person who came on the First Fleet and the value of the crime for which he or she was transported (see Figure 3; data from Watson et al., 2011). We can put these data into a random sampler and then select samples of 10 data values and compare their variation and medians with the actual population measures. For a small sample size taken over several trials we see much variation in the medians of the individual samples (Figure 4). Considered together, 100 of the sample medians have their own median of 26.2 (Figure 5). As the sample size increases to 30 data values, the cluster of medians has less variation (Figure 6) and the median of the medians (29.5) is closer to the population median of 29 (Figure 3).

Some of the students had problems in comparing different samples with the population and making a judgment on how similar or different they were. Part of the reason for this is that outliers changed the X axis for each plot so students had to manipulate the graphs to get comparative scales.

Discussion questions: To what extent might notions of random sampling being fair and representative hold for samples of 10 versus 30? On what are you basing your judgments? What are implications for sampling unknown populations?

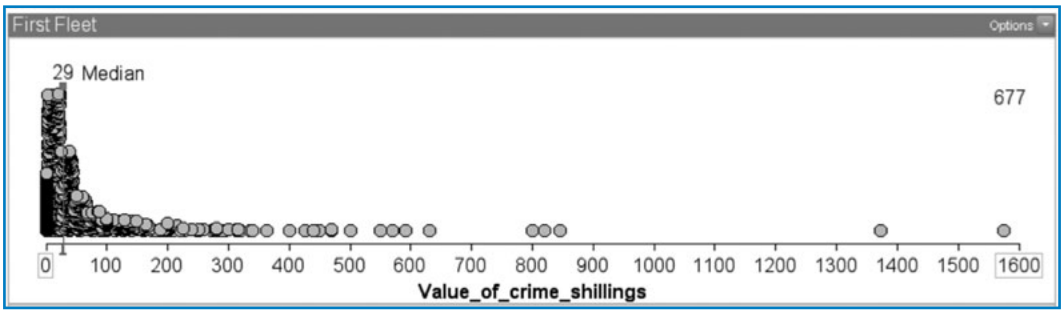


Figure 3. Population from First Fleet with data on Value of Crime in Shillings (one outlier removed).

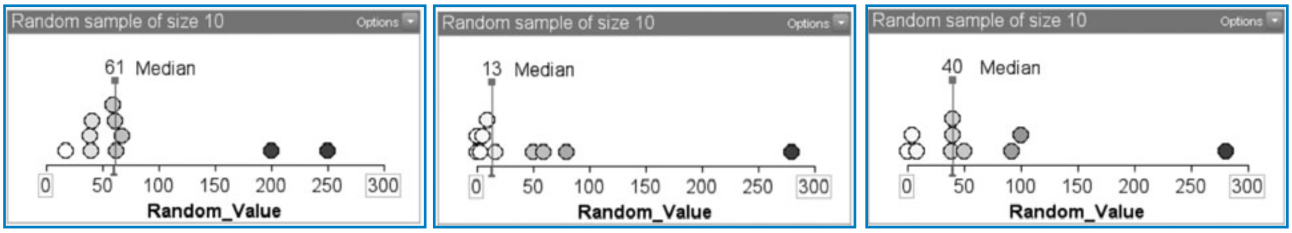


Figure 4. Three random samples of size 10 with medians marked.

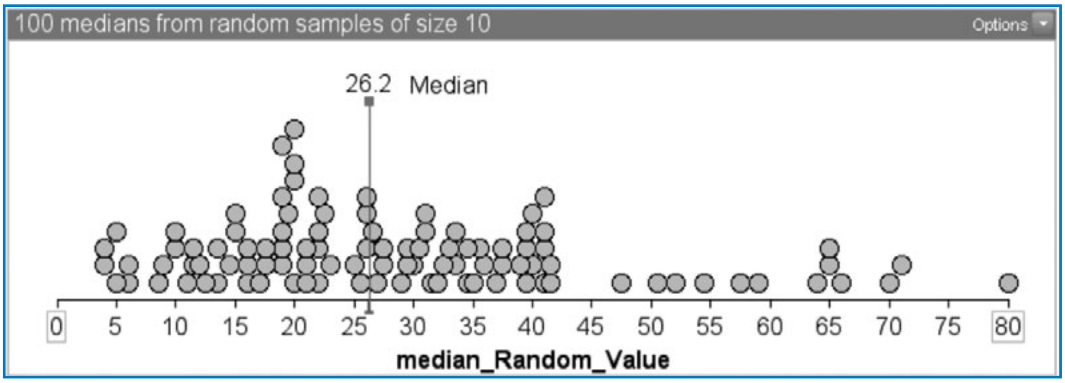


Figure 5. A plot of the medians from 100 random samples of size 10.

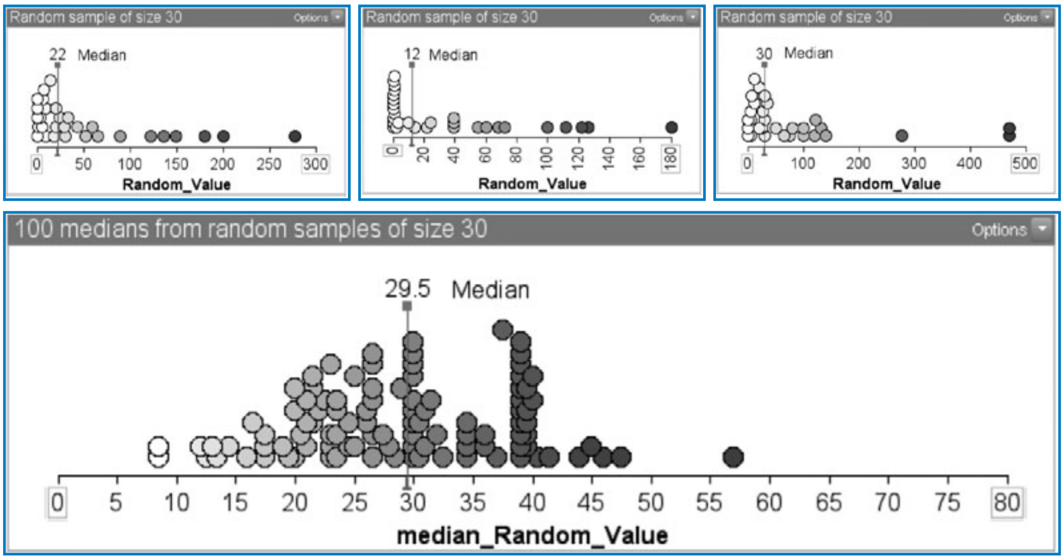


Figure 6. Three random samples of size 30 from First Fleet and medians from 100 such samples.

Random re-allocation of data

Can we use the above ideas to help us think about the dolphin data taken from a small sample of an unknown population? We want to be able to discern if the results mean that the improvement in depression is a caused by swimming with dolphins for our small sample and then be able to generalise to a larger population.

To help us do that we need to imagine that the reverse is true. Despite the logical contextual reasons that swimming with dolphins causes improvement and our expectation that it might be true, we are going to imagine that the dolphin experimental results are merely due to a random extreme event.

Consider each data value as representing one person with information about what *group* he or she is in (dolphin or control) and his or her *result* (improved or not improved). This can be represented by the two-way table in Figure 7. We have 13 people who improved (10 from the dolphin group) and 17 people who did not (5 people from the dolphin group). Now what if the treatment had no effect? We would expect the results of improved/not improved to be randomly spread across the two groups. We can estimate this possibility using a *TinkerPlots* simulation where we put our results into a random generator and then re-assign the results to the two groups (dolphin or control). Details for carrying out the re-allocation with *TinkerPlots* are found in Watson (2013).[1]

We might find that when randomly re-assigned there are only 6 people who 'improve' by swimming with dolphins, then 4 for the next trial, then 8. When we do this for 100 trials we can look at the randomly generated frequency distribution of people who 'improved' by swimming with dolphins (Figure 8).

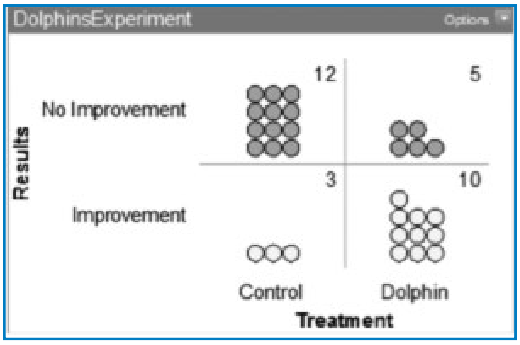


Figure 7. The dolphin experiment two way table with 10 out of 15 people improving after swimming with dolphins.

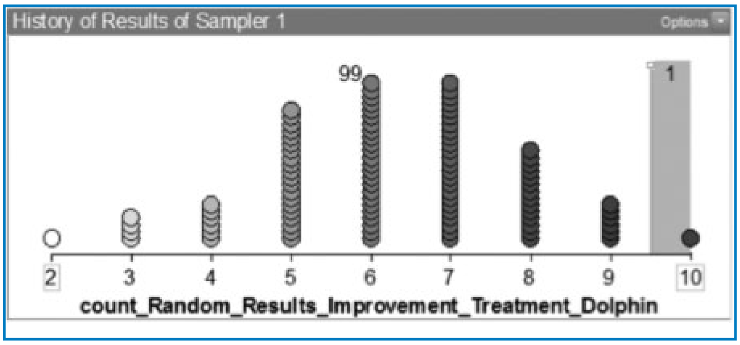


Figure 8. 100 resamples reporting the positive outcomes for improvement after swimming with dolphins.

The random simulation created a spread in values for improvement after swimming with dolphins, and only 1% showed the result that 10 or more people 'improved' when they swam with dolphins. We therefore get a numerical value that suggests the likelihood that the result of the actual experiment was random. We can say, "If we randomly re-allocate the data again and again only about 1% of the time is the randomly generated result the same or more extreme than our experimental result." This gives us high confidence in the actual experimental results for this sample. From this we can take the next step to generalise to a larger population.

In the unit, students also used this process when comparing two different frequency distributions, e.g., the number of meaningful words memorised versus the number of meaningless words memorised—two data sets that they had experimentally generated themselves. This activity was based on suggestions of Shaughnessy, Chance and Kranendonk (2009).

Most students initially found it difficult to imagine that the results of the dolphin treatment could be due to chance, rather than caused by the treatment. They were captured by the conviction that the treatment was logically effective and therefore true.

At the beginning many had difficulty in understanding what random re-allocation was, how it worked and why they were doing it. Some would get the idea, only to lose it later, then regain it. It was conceptually difficult and needed considerable teacher support. With several examples, however, most students developed a visual/intuitive sense of the likelihood that a result could have been a caused versus a random event. It was important to give some examples where the difference was less obvious.

Generalisation

The aim of the random re-allocation process is to help students have more confidence and discernment in generalising from a sample to a larger population due to having a numerical likelihood of it being a random occurrence for a given number of trials. In this step, however, students also need to draw in aspects about the experimental design that they have considered earlier: how certain is it that experimental bias has been mitigated, how was the small sample selected from the population, what population might this be reasonably generalised to? When these

1 This process of random re-allocation is a visually intuitive yet numerically based alternative to a *t*-test approach that some students will meet in later years. It is the technique recommended by Cobb (2007).

factors are considered then students should be able to provide an overall level of confidence in being able to generalise the outcome to a larger population.

Metacognition, imagination and making explicit the thinking processes

The action research project highlighted how important it is to help students create concept maps for the thinking they are doing. It would help if the class developed a glossary and map of the terminology that they were using, particularly notions about randomness, samples and populations. It would be helpful during the course of the unit to spell out the different slang meanings versus statistical meanings, the problems with the meanings as they come up in new conceptual contexts, and questions around these.

It is important to make explicit the sort of thinking that is involved in making effective inferences and discerning judgments. When teachers are drawing on rich contexts in mathematics they may find themselves participating in more general discussions where understanding the context is as important as understanding the mathematics and for the students the lines between the two contexts become blurred. Initially some students drew on other forms of thinking such as scientific thinking in critiquing the experimental design and general reasoning in looking for why an experiment might work, with more focus on the context than on a statistical evaluation of the data. They often felt one dimension of evaluation was enough and did not bring the different types of evaluation together.

In reflecting on the student responses for the dolphin problem it was clear that many of them drew on valuable forms of thinking that need to be made explicit as part of an overall process of making informal inferences. Figure 9 attempts to capture the type of thinking different students were engaged in during the unit and within this context provide a framework for making informal inferences. At every stage of the inference process imagination is required in thinking about the questions posed. Without the questions and imagination to think about alternatives, the process becomes a textbook exercise.

It is important to have metacognitive conversations with students to encourage them to reflect on their thinking and what they are drawing on in coming up with judgments. As they describe the “sources” of their thinking these can be captured, valued, discussed, refined and mapped in a similar way to what is shown in Figure 9 in creating a metacognitive frame for the unit. This then creates visible thinking tools that students can use along with specific statistical thinking such as using the concepts of variation, distribution, expectation, randomness, and informal inference. A key part is helping students overcome limitations in thinking about mathematics being only associated with formulae and calculations, rather, that it requires imagination and flexibility in thinking.

Conclusion

There is considerable research on the difficulties students have in conceptualising individual concepts of probability and statistics (see for example, Bryant & Nunes, 2012; Jones, 2005). The unit of work developed for this action research project was specifically designed to address some of these in order to help students create visual and intuitive understandings of the issues of sampling, randomness and populations. The interweaving of concepts, combined with the technical skills, was challenging for the teachers, students and researchers.

To create meaningful and deep learning experiences for students requires considerable teacher knowledge and skill, being able to weave together concepts with context and metacognitive thinking. In particular, it takes skill in helping

Making statistical judgements about differences in data sets

CONTEXT

1. Hypothesis

Prediction of the generalisation for a wider population based on an experiment to discern different treatments.

2. What are logical reasons that the hypothesis may or may not be true?

3. Experimental validity

- What are the likely biases that could distort the results?
- What strategies can mitigate bias? (e.g., random selection, control group)
- How representative is this sample of the more general population?
- How would you describe this sample size and its adequacy for generalisation?

STATISTICAL INFERENCE

4. Description of the evidence

- Distribution, central tendancy, differences.
- Is there an indication that the difference is significant for these data?

5. Could this have happened by chance?

- How frequently would a *random generation* of these numbers create this experimental result if repeated 100 times in a simulation using *random re-allocation of data?*
- Based on this, how certain are you that this is *caused* result, rather than a *random* one?

6. Generalisability of the experiment

- What is your claim?
- What is the population that you could generalise this claim to?
- What is the certainty of your claim based on the experimental validity and the random re-allocation of data simulation?

7. Implications and recommendations

- Is this result significant enough to recommend any action?
- How might these data be used?
- What questions are still remaining?

Figure 9. Metacognitive frame for making informal inferences for the dolphin context.

students to draw out from diverse activities the central and connecting themes. At the end of the program the lead teacher said she had previously never thought of designing a unit around a key idea such as randomness or inference. Being able to use such big conceptual ideas as a theme for a unit provided a very valuable and interesting experience, certainly deepening her own experience of the nuances within the ideas and how to build more connective conceptual experiences for the students.

For students who go on to study formal statistics, the hope is that the experiences with informal inference in memorable contexts such as the dolphin problem will provide a foundation for appreciating and understanding the formal statistics associated with *t*-tests and *p*-values. For those who do not go on to study formal statistics, it is hoped that they have gained an appreciation of randomness and its usefulness in decision-making, moving beyond seeing random only as haphazard.

References

Bryant, P. & Nunes, T. (2012). *Children's understanding of probability: A literature review.* London: Nuffield Foundation. Retrieved from www.nuffieldfoundation.org

Chiesi, F. & Primi, C. (2009). Recency effects in primary-age children and college students. *International Electronic Journal of Mathematics Education, 4*(3), 259–274.

Cobb, G. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education, 1*(1). Retrieved from http://escholarship.org/uc/item/6hb3k0nz

Jones, G.A. (Ed.). (2005). *Exploring probability in school: Challenges for teaching and learning.* New York: Springer.

Konold, C. & Miller, C.D. (2011). *TinkerPlots: Dynamic data exploration* [computer software, Version 2.0]. Emeryville, CA: Key Curriculum Press.

Moore, D. S. & McCabe, G. P. (1993). *Introduction to the practice of statistics* (2nd ed.). New York: W. H. Freeman.

Shaughnessy, J. M., Chance, B. & Kranendonk, H. (2009). *Focus in high school mathematics reasoning and sense making: Statistics and probability.* Reston, VA: NCTM.

Watson, J. M. (2013). Resampling with TinkerPlots. *Teaching Statistics, 35*(1), 32–36.

Watson, J., Beswick, K., Brown, N., Callingham, R., Muir, T. & Wright, S. (2011). *Digging into Australian data with TinkerPlots.* Melbourne: Objective Learning Materials.

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

Watson, J. M. & Moritz, J. B. (2003). Fairness of dice: A longitudinal study of students' beliefs and strategies for making judgments. *Journal for Research in Mathematics Education, 34*(4), 270–304.