DOING RESEARCH IN STATISTICS EDUCATION: MORE THAN JUST DATA

Jane M Watson
University of Tasmania
Australia

As teachers of statistics we know the fundamental components of statistical enquiry, be it classical or exploratory. When we turn the focus on ourselves as statistics educators, we run the risk of forgetting some of the fundamental principles of good research – principles that are broader than carrying out statistical significance tests. In this talk I want to present some examples of research in statistics education to illustrate the stages and outcomes that contribute to results that have a scholarly impact on the statistics education community. As a single teacher with a good idea on how to teach "confidence intervals," I do not expect anyone to pay much attention to me. If I can, however, place my ideas in the context of others' ideas or research on teaching confidence intervals; conduct a study – maybe a case study or a controlled experimental design – that is valid for considering the issue I want to promote in teaching about confidence intervals; and have my results refereed by peers in the field; then I can expect people to pay attention to me.

Research in statistics education is more broadly based than classical statistics applied to science. So, why do we do research in statistics education? Because research tells us something new and because hopefully research tells us something important. No matter how good the research is, however, how the story is told is what convinces us of the newness and importance. It must also convince us that the research is valid and rigorous in its context.

There are many types of research in statistics education represented at this conference. I only have time to give examples of three. They come from areas I will call theoretical, qualitative and quantitative. These terms reflect the perspectives on dealing with the data collected.

There are many types of content that could be used as a focus for these three kinds of research but I am going to pick three areas in which I have been involved with fellow researchers around the world for the past few years: statistical reasoning, statistical thinking, and statistical literacy [SRTL]. Joan Garfield, who could not be with us at this conference, intended to speak on these themes. She helped organize two forums on SRTL and is currently editing a book on the topic. What we want to do today is to use these three themes to illustrate the three types of research: statistical thinking for a theoretical research perspective, statistical reasoning for a qualitative research perspective, and statistical literacy for a quantitative research perspective. I will have to simplify some details so I encourage you to trace the original sources.

Starting with Statistical Thinking I am reminded of a paraphrase of a HG Wells quote from 1904: "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write" (Castles, 1992, page v). The theoretical research I want to discuss is that of Chris Wild and Maxine Pfannkuch and I want to show that it satisfies the demands of research expected to have an impact on the field (Wild & Pfannkuch, 1999). They are working with ideas and their outcome is a four-dimensional model of statistical thinking.

The researchers were motivated by the following questions: *To think statistically ... what does it mean? How do we tell our students to think statistically? What do practicing students do? What do practicing statisticians do? What do the "experts" say?*

Embarking on a big challenge Wild and Pfannkuch gathered much background from the literature. Examples include elements suggested by the American Statistical Association and David Moore (1997) about the need for data, the importance of data production, the omnipresence of variability, and the measuring and modelling of variability; the thoughts of Ronald Snee (1990) on controlling and reducing variation leading to quality improvement; and suggestions that the "cure" in education is found in students doing projects.

The research "question" or "problem" was to develop a model that included the following aspects: the complex thought processes in solving real-world problems using statistics; a framework for thinking patterns; strategies for problem solving; and integration of statistical elements within the problem solving process. Their methodology, although using the wisdom of those who had spoken in the literature, went further to collect data in interviews with those who had a stake in statistical thinking: 11 students interviewed as they solved textbook-type tasks and critiqued newspaper articles; 5 students who were leaders of project teams solving vaguely indicated problems for clients; and 6 professional statisticians asked about statistical thinking and the projects they had been involved in.

The main outcome of Wild and Pfannkuch's synthesis of the data is a four-dimensional framework for thinking in empirical enquiry: the investigative cycle, types of thinking, the interrogative cycle, and dispositions. These four dimensions are elucidated in figures presented in their paper in the *International Statistical Review*. Some are very complex and require pages to describe. All of us will see ideas with which we are quite familiar in their figures. Why are our names not in print with this model? Because we did not go through all of the stages of the research, including submitting the paper for peer review by the experts in the field – some of the very people who had provided the starting points for the research.

What is the significance of this research? We find others quoting it as a foundation for their own research. It is the starting point for further debate on the nature of statistical thinking. It provides in itself further motivation to collect data, say at the school level, to confirm that the model is appropriate for younger learners as well as those involved in formal empirical enquiry.

The outcome of Wild and Pfannkuch's research is a model for statistical thinking on a large scale. Now we are turning to Statistical Reasoning and will consider research on a smaller scale, specifically related to reasoning about particular topics. The same structure, with slight variations, is the basis of the process. This time we are interested in the development of school students' understanding and motivated by the following questions*: What do students know? How does understanding develop over time?* Potentially interesting topics include chance/probability, average, graphing, inference, and sampling. Jones and his co-workers, for example, have developed frameworks for children's understanding of probability and of statistical reasoning (thinking) (Jones, Langrall, Thornton, & Mogill, 1997; Jones, Thornton, Langrall, Mooney, Perry, & Putt, 2000).

We are going to look at some research on the development of students' understanding of sampling and again follow the process from initial ideas to peer review. The context for research in schools is usually embedded in the curriculum suggested for students. In Australia we find the following:

> The dual notions of sampling and of making inferences about populations, based on samples, are fundamental to prediction and decision making in many aspects of life. Students will need a great many experiences to enable them to understand principles underlying sampling and statistical inference. (Australian Education Council, 1991, page 164).

Further in the United States, the National Council of Teachers of Mathematics (NCTM, 2000) has expectations for students, which are very specific from the elementary level. Upper primary students should "understand what samples are, select appropriate samples from specified groups and draw informal inferences from data collected." High school students should "understand what samples are and recognize the importance of random samples and sample size, and draw inferences and construct and evaluate arguments based on sample data." Grade 9 to 12 students should "acquire intuitive notions of randomness, representativeness, and bias in sampling to enhance their ability to evaluate statistical claims."

The context for this research also includes the previous research in the area. No matter how clever my ideas, they are no use if they repeat what some other researcher did 10 years ago. The purpose of research is to build upon what has been learned by others. This involves searching the literature. In relation to sampling, early work of Tversky and Kahneman (1971 and 1974) involved work with college rather than school students but more recently other research has focused on school students. The researchers include Wagner and Gal (1991), Rubin, Bruce and Tenney (1991), Mokros and Russell (1995), Jacobs (1997), Fischbein and Schnarch (1997), and Schwartz, Goldman, Vye, Barron and The Cognition and Technology Group at Vanderbilt (1998).

In considering student understanding it is also necessary to build on general models of student cognitive development. One such model I have found useful is that of Biggs and Collis

(1982 and 1991). It suggests four levels in the development of reasoning: (a) ikonic, intuitive reasoning; (b) unistructural reasoning: single ideas, contradictions not noticed; (c) multistructural reasoning: sequential use of ideas, recognition of contradictions but no resolution; and (d) relational reasoning: integration of ideas to form a whole, resolution of contradictions.

We also must not forget what statistics itself tells us about sampling, to use as a benchmark in judging outcomes. Two quotes will serve as examples.

> Boswell quotes Samuel Johnson as saying, *You don't have to eat the whole ox to know that the meat is tough*. That is the essential idea of sampling: to gain information about the whole by examining only a part. (Moore, 1991, page 4).
>
> What is the essential nature of a sample? In a word, a sample should be *representative*. This means that, effectively, a sample should be a small-scale replica of the population from which it is selected, in all respects that might affect the study conclusions. (Orr, 1995, page 72).

When it comes to methodology, there are many possibilities with school students. At the moment I am going to discuss a qualitative study based on in-depth interviews. In considering what to ask students in interviews it is also important to build on what has been done before; e.g., for sampling, the work of Kahneman and Tversky (1972) and Konold and Garfield (1992). In being innovative, however, new ideas are likely to be appropriate. In this research, for example, items based on media reports were used, as well as open-ended tasks (that is, questions without a single correct answer). The data set for this study consisted of the responses to interview tasks and questionnaires from 62 students in grades 3, 6 and 9 in government schools in Tasmania.

The content base of the items upon which responses were clustered is summarised briefly as follows: (a) definition of sample, (b) explaining sample in a context (finding the weight of grade 5 children in the state) and suggesting a sample size and method of selection, (c) choosing which of a large or small sample is likely to have an extreme outcome, and (d) recognition of biased samples in two newspaper articles. As an example of the media items, students were asked if they had any criticisms of the following article.

> *ABOUT six in 10 United States high school students say they could get a hand-gun if they wanted one, a third of them within an hour, a survey shows. The poll of 2508 junior and senior high school students in Chicago also found 15 per cent had actually carried a handgun within the past 30 days, with 4% taking one to school.*

The 62 responses were clustered into six categories using techniques suggested by Miles and Huberman (1994), based on *observed* reasoning, the cognitive hierarchy, correctness and curriculum objectives. As in most educational research studies there were some responses that were at times contradictory and hence put in an "equivocal" category. As an educator one hopes this is an indication that the students are in transition to higher levels. The following brief descriptions are used as titles of the categories:

Small Samplers without Selection
Small Samplers with Primitive Random Selection
Small Samplers with Pre-Selection of Results
Equivocal Samplers
Large Samplers with Random/Distributed Selection
Large Samplers Sensitive to Bias

A few examples will be given to illustrate the categories of samplers. *Small Samplers without Selection* (i) may provide examples of samples, such as food products, (ii) may describe a sample as a small bit, or more rarely as a try/test, (iii) agree to a sample size of less than 15, and (iv) suggest no method of selection or an idiosyncratic method. Responses from a grade 3 girl are typical. *Sample?* Like something free, a little packet or something. *How many?* Probably about 10. *How choose?* Teacher might just choose people who have been working well or something.

Small Samplers with Primitive Random Selection are like the previous group but suggest selection "by random" without description, or "choose any," perhaps from different schools.

These students cannot justify their use of "sound good" terms. *Small Samplers with Pre-Selection of Results* are again similar but suggest selection of people by weight, either a spread of fat and skinny, or people of normal weight. The following is an elaborate suggestion by a grade 3 in this category.

> *If they saw a really small person, bring them up and then measure some of the other members of the class with them that they also thought were small. And if there are, the other people are, around their height, they could choose them. And if there was a really tall person, they could measure some of the others with them, and they're around that height. Then you could take those two.*

Equivocal Samplers may indicate indifference about sample size, sometimes based on irrelevant aspects, and either combine small size with appropriate selection methods or partial sensitivity to bias, or large sample size with inappropriate selection methods. In contrast *Large Samplers with Random/Distributed Selection* suggest a sample size of at least 20 or a percentage of the population, and suggest selection based on a random process or distribution by geography. The following response from a grade 6 student illustrates typical selection methods in this category: *Say take a kid from each school. Take some – just pick a kid from random order. Look up on the computer; don't even know what the person looks like or anything. Pick that person.* Students in this group, however, do not detect bias in the newspaper articles. When asked for criticisms of the article on guns in the United States, this student replied, *I think they feel that they're safe with a hand gun and they aren't stupid with it; it's their personal business*.

Large Samplers Sensitive to Bias possess the characteristics we would want in dealing with samples, including questioning bias. They provide examples of samples, sometimes involving surveying; describe a sample as both a small bit, and a try/test; may refer to terms "average" or "representative"; suggest a sample size of at least 20 or a percentage of the population; suggest selection based on a random process or distribution by geography; express concern for selection of samples to avoid bias; and identify biased samples in newspaper articles reporting on results of surveys.

In terms of development of understanding over the years of schooling, we found that all grade 3 students responded in one of the Small Sampler categories and all of the grade 9 students were in the top three categories (Equivocal or Large Samplers). Grade 6 ranged over the top five categories but were mainly found in the three categories centered around Equivocal.

Several recommendations arise from this research for educators and curriculum developers. These include the need for emphasis to be placed on the transition from an out-of-school (homogeneous) to an in-school (heterogeneous) meaning of "sample", sample size for representativeness, recognition of bias, and use of media examples based on social settings.

Also in relation to the recommendations of others we support two in particular. One is the plea of Joram, Resnick and Gabriele (1995, page 359) to use the media where "the goal is explicitly to comprehend and interpret information in a passage rather than to solve a specific problem." The other is a comment of Derry, Levin, Osana and Jones (1998, page 190) that "statistical reasoning does not [always] necessitate computation, but always involves interpreting and reasoning about real-world problems with conceptual structures."

The next step in the process is to submit the whole story to peer review and hope that your colleagues agree you have made a contribution to the field (Watson & Moritz, 2000a and in press). Most avid researchers do not stop here, however. Their initial research suggests more questions that are the starting points for further research. For me some of these questions include the following*: What happens when students work collaboratively? Does it produce high level outcomes?* (Chick & Watson, 2001; Watson & Chick, 2001a and 2001b*). What about cognitive conflict? Does it aid understanding?* (Watson, 2002; Watson & Moritz, 2001). *Can we teach to improve understanding, say with respect to variation?* (Watson & Kelly, 2002).

Third I want to suggest briefly two examples of some quantitative research based on a large sample related to the third theme, that of Statistical Literacy. The first is a complement to the qualitative research just described. Although Iddo Gal has had much to say on the subject of statistical literacy, Katherine Wallman (1993) has a succinct summary for the purposes of this research:

> *Statistical Literacy* is the ability to understand and critically evaluate statistical results that permeate our daily lives – coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions. (page 1).

Content is obviously very important in this definition. If the overall aim of statistics educators is to produce people to satisfy Katherine Wallman's definition, we need some standards by which we can judge progress through the years of schooling. A model designed to describe the hierarchical objectives of statistical literacy is useful at this point (Watson, 1997). The three tiers of the hierarchy are (1) understanding the basic terminology, (2) understanding terminology in social contexts, and (3) developing the ability to question statistical claims which are made without proper justification.

This research followed a similar overall structure to the preceding study but was based on surveys of over 3000 students in grades 3 to 11, with some longitudinal data collected after two or four years (Watson & Moritz, 2000b). Using items reflecting the demands of the tiers of statistical literacy, clustering assisted by language analysis software (Qualitative Solutions and Research, 1995) confirmed structural patterns that could be identified as fitting the levels of the cognitive development model (Biggs & Collis, 1982 and 1991), plus a further consolidation level where responses identified bias. At grade 3, for example, 48% of students gave inappropriate responses to what a sample was (not reaching Tier 1), whereas 3% gave a complete relational description of sample, with the others providing unistructural or multistructural descriptions. By grade 11, only 6% could not reach Tier 1, whereas 46% could identify bias in one context and 20% could identify it in two contexts (Tier 3). Recommendations were similar to before and confirm recent work of Gal (e.g., 2002) from a quantitative perspective.

The second quantitative example takes us into suggestions for *further* research, which has not yet completed the entire process to the peer review stage. It would be useful to describe a global underlying statistical literacy variable that would assist in assessing improved performance across the tiers of Statistical Literacy using surveys. Step 1 in the process is the large scale analysis of 3000 responses from students, in grades 3 to 11 to 80 items covering probability, average, sampling, graphs, tables, variation and inference. Of interest educationally are the mathematical/statistical skills involved, engagement with the contexts, and the structure of questions asked. For this task a different quantitative method called Rasch analysis is used (1980). It is a likelihood model that simultaneously considers student ability and item difficulty to produce a variable map of student ability versus item difficulty. The procedure assumes unidimensionality, which is judged by goodness of fit. Already the method has been used by Reading (2002) to confirm hierarchical aspects of statistical understanding in her data. I expect to find aspects associated with the three tiers of statistical literacy based on coding of responses reflecting a cognitive hierarchy and correctness. It is early yet, but this type of analysis offers promise in developing instruments to measure statistical literacy understanding on a single dimension. *Will such a model hold up under scrutiny with fewer items, or trials in other countries? Will it satisfy our peers under review?* It will be interesting to see. It is the responsibility of researchers, however, to take risks and make suggestions for the future.

I want to close with a bit of serendipity, which illustrates why I keep going with my research into children's understanding of statistical concepts. Sometimes the real world impinges on research in very surprising but helpful ways. When we began our interview research, we wanted to explore students' understanding of the mean represented as a decimal. At this time there was an advertisement for Ford motor cars on television, which claimed, "The average young family has 2.3 children." Then a young boy appeared with a huge ".3" on his sweatshirt, saying "Yea and I'm the Point 3." The punch line of the advertisement was "small family cars often forget poor point 3", stressing the size advantage of the new Ford Laser. We did not refer to the advertisement in our interviews but asked students in grades 3 to 9 what it meant to say an average young family had 2.3 children. A full range of responses was given, from no understanding to a well-justified explanation. Of some interest, however were the responses apparently influenced by the media and other real-life situations. A grade 5 replied, *Probably because the 2 means that there are 2 children and the .3 is probably a small child. It doesn't* *equal out to 3 children.* A grade 6 expressed exceeding cognitive conflict: *Well, a mum might have 2 children or something, and she might be pregnant ... [but] on the ad they didn't do it like that.* And perhaps surprisingly, a grade 9 offered a complex explanation:

> *Um ... because the 2 is like the older children, which they could be fully grown or my age or whatever, and the 3 is a child that's growing up to be an older child so that, like, say the kid is 3 now, once it turns to be 10 it will get to be 1, so they will have 3 children ... sort of thing.*

When the media environment in which students exist exerts such an influence on their statistical literacy skills, we know we have a very important job to do as statistics educators. Valid research is an important part of that job.

REFERENCES

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

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. Academic Press: New York.

Biggs, J.B., & Collis, K.F. (1991). Multimodal learning and the quality of intelligent behaviour. In H.A.H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 57-76). Hillsdale, N. J.: Lawrence Erlbaum.

Castles, I. (1992). *Surviving statistics: A user's guide to the basics.* Canberra: Australian Bureau of Statistics.

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

Derry, S.J., Levin, J.R., Osana, H.P., & Jones, M.S. (1998). Developing middle-school students′ statistical reasoning abilities through simulation gaming. In S.P. Lajoie (Ed.), *Reflections on statistics: Learning, teaching and assessment in grades K-12* (pp. 175-195). Mahwah, NJ: Laurence Erlbaum.

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.

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.

Jones, G.A., Langrall, C.W., Thornton, C.A., & Mogill, A.T. (1997). A framework for assessing young children's thinking in probability. *Educational Studies in Mathematics, 32*, 101-125.

Jones, G.A., Thornton, C.A., Langrall, C.W., Mooney, E.S., Perry, B., & Putt, I.J. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning, 2*, 269-307.

Joram, E., Resnick, L.B., & Gabriele, A.J. (1995). Numeracy as cultural practice: An examination of numbers in magazines for children, teenagers and adults. *Journal for Research in Mathematics Education, 26*, 346-361.

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

Konold, C., & Garfield, J. (1992). *Statistical reasoning assessment. Part 1: Intuitive Thinking*. (Draft document.) Scientific Reasoning Research Institute, University of Massachusetts, Amherst, MA.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook.* (2nd ed.). Thousand Oaks, CA:Sage Publications.

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

Moore, D.S. (1991*). Statistics: Concepts and controversies*. (3rd ed.). New York: Freeman.

Moore, D. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*, 123-165.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Orr, D. B. (1995). *Fundamentals of applied statistics and surveys*. New York: Chapman and Hall.

Qualitative Solutions and Research. (1995). *Non-numerical unstructured data · indexing searching and theorizing (NUD·IST) v3.0.5*. Melbourne: LaTrobe University (computer program).

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press (original work published 1960).

Reading, C. (2002). *Profile for statistical understanding*. Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

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: International Statistical Institute.

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.

Snee, R. (1990). Statistical thinking and its contribution to quality. *The American Statistician, 44*, 116-121.

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

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Wagner, D.A., & Gal, I. (1991). *Project STARC: Acquisition of statistical reasoning in children*. (Annual Report: Year 1, NSF Grant No. MDR90-50006). Philadelphia, PA: Literacy Research Center, University of Pennsylvania.

Wallman, K.K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88*, No. 421, 1-8.

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. (2002). *Creating cognitive conflict in a controlled research setting: Sampling.* Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

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

Watson, J.M., & Chick, H.L. (2001b). Factors influencing the outcomes of collaborative mathematics problem solving—An introduction. *Mathematical Thinking and Learning, 3*(2&3), 125-173.

Watson, J.M., & Kelly, B.A. (2002). *Can grade 3 students learn about variation?* Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

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

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

Watson, J.M., & Mortiz, J.B. (2001). The role of cognitive conflict in developing students' understanding of chance measurement. In J. Bobis, B. Perry, & M. Mitchelmore (Eds.) *Numeracy and beyond* (Proceedings of the 24th Annual Conference of the Mathematics Education Research Group of Australasia, Vol. 2, pp. 523-530). Sydney, NSW: MERGA.

Watson, J.M., & Moritz, J.B. (in press). Developing concepts of sampling for statistical literacy. In J. Sowder & B. Schappelle (Eds.), *Lessons learned from research*. Reston, VA: National Council of Teachers of Mathematics.

Wild, C.J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*, 223-265.