DISCERNING STUDENTS' STATISTICAL THINKING: A RESEARCHER'S PERSPECTIVE

James Baglin
School of Mathematical and Geospatial Sciences
RMIT University, Melbourne, Australia
james.baglin@rmit.edu.au

In the most comprehensive treatment of the topic to date, Wild and Pfannkuch (1999) propose a structural overview of the various domains of what constitutes statistical thinking. This paradigm provides a useful framework for approaching statistical thinking assessment. However, translating this model into assessments that are practical, reliable and valid for the purpose of statistics education research remains a challenge. This paper discusses the challenge of assessing statistical thinking and the development of a preliminary research-based assessment task based on Wild and Pfannkuch's paradigm. Data collected from a large introductory statistics course where students completed the preliminary task are critically evaluated. Suggestions for future improvements to the task and ideas for alternative methods are raised. Statistical thinking may very well prove to be as difficult to assess as it is to define, but without further research, our understanding of how people learn to think statistically will be limited.

STATISTICAL THINKING

Many statistics researchers have contributed to our knowledge of statistics education outcomes. Statistical literacy, reasoning and thinking are common terms used interchangeable or distinctly to refer to different layers of statistical understanding (e.g. Garfield, delMas, & Zieffler, 2010). However, to this day, there remains no single agreed upon definition or taxonomy of statistical learning outcomes (Ben-Zvi & Garfield, 2005). Statistical thinking, once defined shrewdly as an understanding of what a statistician does (Chance, 2002), has presented instructors and researchers alike with a unique challenge. The work of Chambers (1993) and later Cameron (2009) are a useful place to begin the elaboration of "what a statistician does". Chambers and Cameron proposed five characteristics of effective consulting statisticians, which included the following: i) preparing data, including planning, collection, organization and validation, ii) analyzing data, by models or other summaries, iii) presenting data in written, graphical or other form, iv) formulating a problem so that it can be addressed through statistical means, and v) carrying out research to develop new statistical methods. Beside the final point, which relates specifically to practicing statisticians, these capabilities share much in common with the desired learning outcomes of modern introductory statistics courses (e.g. American Statistical Association, 2005).

Building on this notion, the seminal work of Wild and Pfannkuch (1999), and later Pfannkuch and Wild (2000, 2005), provided statistics educators with the most comprehensive treatment of what it means to think statistically. According to Wild and Pfannkuch the core of statistical thinking is problem solving. Wild and Pfannkuch reasoned that statisticians approach problems using similar investigative, thinking, reasoning and dispositional approaches. Fundamental to their paradigm was the idea of data investigation as a process. Wild and Pfannkuch (1999) adopted the PPDAC (Problem, Plan, Data, Analysis, Conclusions) cycle, first proposed by MacKay and Oldford (1994, as cited in Wild & Pfannkuch,1999). The framework also proposed five types of fundamental thinking including recognizing the need for data, transnumeration, consideration of variation, reasoning with statistical models, and integrating the statistical and contextual (see Table 1 for definitions). Wild and Pfannkuch's paradigm did not provide a "sound bite" definition of statistical thinking, nor was that their aim. Instead, Wild and Pfannkuch provide a useful framework for conceptualizing a complex, broad, holistic and multidimensional way of thinking. Therein lays the challenge for instructors and researchers. How do we reliably and validly discern statistical thinking in our students given its multifaceted nature?

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

ASSESSING STATISTICAL THINKING

Statistical thinking appears to be as difficult to assess as it is to define. Garfield and delMas (2010) provide an overview of the ARTIST (Assessment Resource Tool for Improving Statistical Thinking) Project, which developed enhanced traditional assessment items (multiple-choice and short-answer questions) recommended for the assessment of statistical thinking. These ARTIST resources are examples for how traditional methods can be enhanced (Wild, Triggs, & Pfannkuch, 1997), however, as Chance (2002) observed "evidence of statistical thinking lies in what students do spontaneously, without prompting or cue from the instructor" (p. 130). Open-ended methods would appear more suitable as they allow students to demonstrate their knowledge by constructing their answers (Watson, 1997). However, open-ended methods of assessment require a strong conceptual framework to guide marking, which is often lacking due to the nebulous nature of statistical thinking. Assessment of problem-based learning and project-based learning outcomes are well aligned to Wild and Pfannkuch's paradigm as they promote a holistic engagement with the entire data investigation cycle (MacGillivray & Pereira-Mendoza, 2011). The use of statistical data investigation projects have been reported extensively in the literature (e.g. Griffiths & Sheppard, 2010; Holmes, 1997; Potthast, 1999; Smith, 1998), but it was the work of Marriott, Davies and Gibson (2009) that highlighted the challenge of assessing statistical problem solving.

Marriott et al. (2009) reported on the problem-based learning reform of the statistics component of the UK primary and secondary school mathematics curriculum. Marriott et al. discussed the extensive re-development of assessment practices required. As the authors explained, the open-ended nature of problem-based learning creates variability in solutions and, as such, would require additional grading time. The authors developed careful marking schemes that considered each step of the problem-solving approach, but still allowed the holistic nature of the data investigative process to be considered. Marriott et al.'s work demonstrates that practical and valid methods for assessing statistical thinking as problem solving are beginning to emerge, but more work is needed to inform the development of assessment tools that can be used for research related purposes. In line with these findings, this paper reports the development of a preliminary assessment task for measuring statistical thinking in line with Wild and Pfannkuch's paradigm.

THE TASK

A preliminary statistical thinking task was developed to measure indicators of students' statistical thinking at the end of an introductory course. If the task was to be suitable for research related purposes, it had to be brief, but insightful. Due to the multidimensional nature of statistical thinking, it was unreasonable to expect the tasks to capture indicators of all elements of statistical thinking according to Wild and Pfannkuch. Therefore, the "investigative" and "types of thinking" dimensions were targeted due to their perceived ease of translation to an assessment tool and significance within the paradigm. The task used an open-ended, short-answer format so students were required to construct their answers. This approach was expected to provide a richer insight into students' thinking, as opposed to a multiple-choice format that would run the risk of students getting the right answer for the wrong reason (Jolliffe, 2010). The task provided students with the following two research problems (the P of PPDAC):

1. Observational Study: Suppose you need to conduct an observational/correlational study that will determine if there is statistical evidence of an association/relationship between eating a diet high in protein and body fat percentage. Explain how you would design, conduct and analyse the results of your study by addressing each of the following points:
2. Experimental Study: Suppose you need to conduct an experiment that will determine if caffeine consumption prior to a lecture helps to improve university students' attention. Explain how you would design, conduct and analyse the results of your experiment by addressing each of the following points:

Students were then asked to address the following six sub-tasks, aligned to the PPDAC cycle, for each scenario.

I. Explain how you would obtain a sample for your study. (2nd P of PPDAC)
II. Explain what data you need to gather to answer the research question and how you would go about obtaining it (2nd P and D of PPDAC)
III. Based on the data that you proposed to gather in II, explain how you would plan to summarise and present the results of the study. (A of PPDAC)
IV. Which statistical test would you use to perform hypothesis testing based on the data that you proposed to gather in II and summarise in III? Justify your choice of test. (A of PPDAC)
V. In your own words, explain why it is important to perform hypothesis testing for this study. (A and C of PPDAC)
VI. Assume at the end of the study you find evidence of an association/relationship. Explain what you expect your summary data and hypothesis testing results to look like. (C of PPDAC)

Table 1 maps each sub-task to Wild and Pfannkuch's types of thinking dimension. The reason both an observational and experimental scenario were presented was to assess students' statistical thinking in different contexts and research situations. Similar to Marriott et al. (2009), a grading scheme was developed for each scenario to help standardise and identify the main indicators of statistical thinking in each task. Each sub-task for each scenario was graded on a scale ranging from High (3 points) to Poor (0), reflecting the degree to which the students' exhibited statistical thinking about the scenario and sub-task presented. The grading scheme, too large to report here, considered the students' elicitation of the following indicators for each sub-task: i) population, sample, sampling, representative sample, statistical power analysis, ii) study design, random allocation vs. observation, operationalization and measurement of variables, manipulation of variables, iii) linking types of data collected with correct descriptive and graphical summaries, iv) appropriate statistical models/method selection based on data collected, assumptions, v) sampling variability, uncertainty, inference from samples, null hypothesis significance testing, vi) hypothesised statistical effects, awareness of visual statistical cues, statistical significance.

Table 1. *Mapping Wild and Pfannkuch's (1999) Types of Thinking Dimension to the Statistical Thinking Sub-tasks*

Type of Thinking	Sub-task
1. The recognition of the need for data	i, ii
2. Transnumeration - Identifying and transforming appropriate data into representations of a model that leads to understanding. This occurs at multiple stages – obtaining data to answer a research question and transforming data (e.g. descriptive statistics, plots) to convey meaning and understanding (Pfannkuch & Wild, 2000).	ii, iii
3. Consideration of variation - Knowledge and understanding comes with uncertainty due to the omnipresence of variation, e.g. sampling variability.	i, ii, v
4. Reasoning with statistical models - Understanding of statistical models, how they relate to research design, and how they contribute to understanding	iii, iv, v
5. Integrating the statistical and contextual - Integrating and interpreting statistics within the context of the problem	iii, iv, vi

IMPLEMENTATION AND FINDINGS

Following ethics approval, the statistical thinking task was piloted in a large undergraduate introductory statistics course for science students at end of the semester during a regular tutorial. The course ran for 12 weeks covering the design of experiments, ethics, exploratory data analysis, probability and statistical inference. Contact hours included three hours of lectures and two hours of tutorials per week. Course assessment involved weekly quizzes (15%), a paper review (15%), a major course project (20%) and exam (50%). The learning outcome of the course project was to develop students' statistical thinking by engaging them in the PPDAC cycle. For the project,

students had to work individually or in groups of up to three to design, implement and report the findings of an open-ended statistical investigation using the online virtual environment known as the *Island* (for full details of the projects see Bulmer & Haladyn, 2011). The project mark was split between a proposal (5%) and conference style abstract (15%).

The tasks were completed online during the final tutorials of the semester. Students were given a participation mark for completing the task. There were a total of 574 students enrolled in the course that semester, of which 356 (62%) consented to have their project marks and task results recorded for research purposes. All task attempts by the consenting students were graded by the author of the paper. The other attempts from the non-consenting students were marked by tutors.

Exploratory factor analysis was conducted to examine the underlying structure of all 12 sub-tasks. There were 288 participants who provided valid responses to all tasks from both scenarios. Exploratory factor analysis was conducted using *FACTOR* (Lorenzo-Seva & Ferrando, 2006), and, due to the ordinal nature of the grading scale, was based on analysing the polychoric correlation matrix. The results of a parallel analysis using minimum rank factor analysis (Timmerman & Lorenzo-Seva, 2011) supported a unidimensional solution that explained 56% of the common variance (Cronbach's $\alpha = .87$, see Table 2).

Initial validation of the task was considered by correlating the total task score with project marks. As project marks were highly left skewed, a nonparametric Spearman rank correlation, ρ, was calculated between the total statistical thinking task score and students' final project mark given out of 20. Note, $N = 285$ due to some students not submitting projects. A 95% bias-corrected and accelerated (BCa) bootstrapped confidence interval (Efron, 1987) for ρ was also computed. The results estimated Spearman's $\rho = .27$, 95% Bootstrap BCa CI (.16, .39). The total task score was positively and significantly correlated with project marks.

Table 2. *Exploratory Factor Analysis Solution of the Statistical Thinking Sub-Tasks*

Task	*M*	*SD*	Loading	Communality
Observational I	2.46	0.74	0.53	0.72
Observational II	2.68	0.92	0.62	0.76
Observational III	2.52	0.96	0.69	0.74
Observational IV	2.50	0.99	0.69	0.69
Observational V	1.85	0.84	0.61	0.80
Observational VI	2.54	0.91	0.71	0.71
Experimental I	2.43	0.69	0.47	0.59
Experimental II	2.69	0.73	0.63	0.75
Experimental III	2.44	0.94	0.72	0.84
Experimental IV	2.66	0.99	0.76	0.86
Experimental V	1.92	0.84	0.66	0.98
Experimental VI	2.72	0.95	0.74	0.84
Total[a]	29.4	6.8		

[a] Mean score out of 36 after summing all sub-tasks, $N = 285$.

DISCUSSION AND CONCLUSIONS

The multifaceted nature of statistical thinking presents an assessment challenge to anyone interested in understanding its' development. Wild and Pfannkuch's paradigm captures the complexity of statistical thinking, devoid of a clean definition, but still provides insight into where indicators of statistical thinking will lie. At the core of the range of possible assessment targets is the concept of statistical thinking as a process of problem solving. This preliminary work has developed a statistical thinking task based on a known paradigm with the goal to inform the future development of valid and reliable tools for assessing and researching the development of statistical thinking. The task focused on discerning students' ability to think statistically according to the first two dimensions of Wild and Pfannkuch's model, namely, the data investigative cycle and types of

thinking. The short-answer format ensured that students constructed their responses in order to gain a deeper understanding into their thinking ability. The task is not revolutionary, and may in fact reflect similar assessment strategies already used by many instructors. However, this task is the first, to the best of the author's knowledge, to align and validate a statistical thinking assessment tool to the Wild and Pfannkuch paradigm.

The tasks were piloted on a sample of undergraduate science students at the end of the semester after completing a major course project that engaged them in project-based learning of the entire data investigative process. Exploratory factor analysis suggested a unidimensional solution to the tasks with high internal consistency. This may seem surprising under the premise that statistical thinking is multidimensional, however, this can be explained by the fact that each sub-task tapped into multiple thinking types and that the task was designed to provide an overall measure of statistical thinking as a problem solving cycle. It was not designed with the sensitivity to detect the nuance of the different, but related, thinking types. The task was also found to demonstrate a positive, but somewhat low, correlation with project marks, providing initial evidence of the task's criterion validity.

There are a few limitations to this preliminary work. The analysis was based on only 50% of students in the course and the tasks were only formative in assessment. Response and task apathy bias are possible. The tasks were designed for science students, and therefore won't be suited to all disciplines. However, the tasks can be readily adapted to different situations, for example, by changing the research scenarios. The open-ended and short-answer response format required moderate marking time, but was aided by a clear grading scheme as recommend by Marriott et al. (2009). The short-answer format favoured strong written communication skills. While this format was a practical decision, other, albeit less practical, methods may provide further insight, for example viva-voce or possibly drawing. Enhanced multi-choice formats should not be ruled out, but careful development and validation must be undertaken. The task, or future tasks to emerge from this work, will require further validation. The inter-rater and test-retest reliability of the task requires consideration. Future work should concentrate on measuring the predictive ability of the task by comparing task performance between students and those considered to be experienced statistical thinkers. The task must also be applied and validated in diverse populations and across different disciplines that apply statistics.

The reliable and valid assessment of statistical thinking is of utmost importance to statistics educators and researchers. Wild and Pfannkuch have provided a valuable framework to initiate the advancement of the assessment of statistical thinking beyond the observation that "We know it when we see it", (Wild & Pfannkuch, 1999, p. 223). Statistical thinking may very well prove to be as difficult to assess as it is to define, but without further research, our understanding of the development of statistical thinking will continue to challenge the field.

ACKNOWLEDGEMENTS

This study obtained institutional ethics approval from the University of Queensland on the 25th of January 2012 (Project No. 2011001393).

REFERENCES

American Statistical Association. (2005). *Guidelines for assessment and instruction in statistics education: College report*. Alexandria, VA: Author. Retrieved from http://www.amstat.org/Education/gaise/GaiseCollege_Full.pdf

Ben-Zvi, D., & Garfield, J. B. (2005). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3–15). New York: Kluwer Academic Publishers.

Bulmer, M., & Haladyn, J. K. (2011). Life on an Island: A simulated population to support student projects in statistics. *Technology Innovations in Statistics Education, 5*. Retrieved from http://escholarship.org/uc/item/2q0740hv

Cameron, M. (2009). Training statisticians for a research organisation. *Proceedings of the 57th Session of the International Statistical Institute*. Durban, South Africa. Retrieved from http://isi.cbs.nl/iamamember/CD8-Durban2009/A5 Docs/0276.pdf

Chambers, J. M. (1993). Greater or lesser statistics: A choice for future research. *Statistics and Computing*, *3*, 182–184.

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*. Retrieved from http://www.amstat.org/publications/jse/v10n3/chance.html

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*, 171–185.

Garfield, J. B., & delMas, R. (2010). A web site that provides resources for assessing students' statistical literacy, reasoning and thinking. *Teaching Statistics*, *32*, 2–7.

Garfield, J. B., delMas, R., & Zieffler, A. (2010). Assessing important learning outcomes in introductory tertiary statistics courses. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 75–86). Chichester, WS: John Wiley & Sons.

Griffiths, P., & Sheppard, Z. (2010). Assessing statistical thinking and data presentation skills through the use of a poster assignment with real-world data. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 47–56). Chichester, WS: John Wiley & Sons.

Holmes, P. (1997). Assessing project work by external examiners. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 153–164). Amsterdam: IOS Press.

Jolliffe, F. (2010). Assessing statistical thinking. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 71–74). Chichester, UK: John Wiley & Sons.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*, 88–91.

MacGillivray, H., & Pereira-Mendoza, L. (2011). Teaching statistical thinking through investigative projects. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education: A Joint ICM/IASE Study: The 18th ICMI Study* (pp. 109–120). New York, NY: Springer Science+Business Media B. V.

Marriott, J., Davies, N., & Gibson, L. (2009). Teaching, learning and assessing statistical problem solving. *Journal of Statistics Education*, *17*. Retrieved from http://www.amstat.org/publications/jse/v17n1/marriott.pdf

Pfannkuch, M., & Wild, C. (2000). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. *Statistical Science*, *15*, 132–152. Retrieved from http://www.jstor.org/stable/2676728

Pfannkuch, M., & Wild, C. (2005). Towards an understanding of statistical thinking. In D. Ben-Zvi & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17–46). New York: Kluwer Academic Publishers.

Potthast, M. J. (1999). Outcomes of using small-group cooperative learning experiences in introductory statistics courses. *College Student Journal*, *33*, 34–42.

Smith, G. (1998). Learning statistics by doing statistics. *Journal of Statistics Education, 6*. Retrieved from http://www.amstat.org/publications/jse/v6n3/smith.html

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*, 209–20.

Watson, J. M. (1997). Assessing statistical thinking using the media. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107–121). Amsterdam: IOS Press. Retrieved from http://www.stat.auckland.ac.nz/~iase/publications/assessbk/chapter09.pdf

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