NOVICE STUDENTS' CONCEPTUAL KNOWLEDGE OF STATISTICAL HYPOTHESIS TESTING

Anne M. Williams
Centre for Mathematics and Science Education
Queensland University of Technology
<am.williams@qut.edu.au>

Examination of the statistical literature shows that consensus on definition, terminology, and interpretation of some hypothesis testing concepts is elusive. This makes hypothesis testing a difficult topic to teach and learn. This paper reports on the results of a study of novice students' conceptual knowledge of four hypothesis testing concepts through talking aloud and interview methods. While some students seemed to have a reasonable understanding of some concepts, many students seemed to have more limited understanding. The study explores students' faulty conceptual knowledge.

It is widely recognised that statistics, and in particular hypothesis testing, is a difficult subject to teach and learn (Garfield & Ahlgren, 1988; Hawkins, 1991). In fact, the statistical literature shows evidence of misconceptions at all ages and levels of expertise. It is believed that the nature of the subject itself presents problems because *"the important concepts of statistics are quintessentially abstract* [italics in original]" and open to interpretation (Watts, 1991, p. 290), and are "unlike anything the student has thought of before" (Garfield & Ahlgren, 1986, p. 271). Moreover, it has been suggested that past teaching practices have contributed to difficulties because of an overemphasis on formulae and routine techniques (Ehrenberg, 1990), which encourage rote learning and computation, rather than reflection and interpretation (Hawkins, Jolliffe, & Glickman, 1992).

A study of the literature on particular concepts in hypothesis testing demonstrates that many inconsistencies associated with definitions, terminology, and interpretations exist. With respect to definitions, Truran (1998) observed a lack of consistency and comprehensiveness in his study of text book definitions of the null hypothesis. Freund and Perles (1993) found four different definitions of p-value. With respect to terminology, level of significance, for example, has been termed significance level or *alpha,* and is represented by the symbol a or (less commonly) by the symbol $P_{critical}$ (see Thompson, 1994). P-value has been referred to as "significance probability," "probability level" (Huberty, 1985), "*p*" (Carver, 1978), "$P_{calculated}$" (Thompson, 1994), "prob-value," "tail probability," "P," "P-value," and "descriptive significance level" (Freund & Perles, 1993). Interpretation of concepts presents an even greater problem. For example, there are many stances regarding the issue of "acceptance" of an hypothesis (e.g., Frick, 1996; Hagen, 1997; Serlin, 1993). Gigerenzer (1993) reported finding nine different interpretations of the level of significance in a single textbook. Misinterpretations of this concept were attributed to confusion of the conditional probabilities associated with it (Falk, 1986). In particular, Carver (1978) criticised the interpretation of p-value as the probability of the sample result occurring by chance, the "Odds-Against-Chance Fantasy". In addition, misinterpretations of the significance concept have included perceiving it as a way to establish validity or reliability (Menon, 1993), a means of providing confidence (Chandler, 1970) or importance (West, 1990), or as the goal of research (Carver, 1993).

In summary, the literature highlights the range and complexities of problems associated with defining, representing, and interpreting hypothesis testing concepts, but provides relatively little empirical research on students' understanding of these concepts. It would be expected that students have some knowledge of the issues discussed in the literature in order to gain conceptual understanding of hypothesis testing. This paper reports the results of a study of novice students' conceptual knowledge of hypothesis testing concepts after the completion of a semester's study of statistics.

THE STUDY

Eighteen volunteer students from a large class enrolled in a university-level introductory statistics subject were interviewed after their final examination in the subject. The aim was to explore their knowledge of elementary hypothesis testing (one- and two-sample z and t tests only), a major component of their subject. These students were asked to talk aloud while they worked through three tasks, a Concept Mapping Task and two typical Hypothesis Testing tasks. In the Concept Mapping Task, students were provided with a number of labels marked with concept names associated with hypothesis testing. They were asked to arrange them in such a way as to show the relationships between them, and were encouraged to label the connecting links. *Typical* Hypothesis Testing tasks meant that summary data was provided, numerical calculations could be performed, and data exploration was limited. After the completion of each task, a semistructured interview was conducted in order to encourage the expansion and exploration of ideas touched on during the performance of the tasks themselves.

This paper reports on the students' conceptual knowledge of hypothesis testing evident from the talking aloud tasks and the individual interviews. Given the large number of concepts associated with hypothesis testing, the results reported here concentrate on four major concepts, namely hypothesis, significance level, p-value, and significance. Conceptual knowledge is defined as (a) the knowledge about concepts (revealed in definitions, examples/non-examples, and discussion of issues, features, properties, uses, or limitations associated with concepts); and (b) the knowledge about relationships between and among the concepts. It is assumed that the greater the knowledge about the concepts and their integration with others, and the greater the number and strength of these relationships, then the better the level of conceptual knowledge is, and the more accessible it is for higher-level thinking (Chi & Koeske, 1983). As students found difficulty applying linking words between the concept labels, it was the students' dialogue, rather than the concept maps, that was more useful in revealing their conceptual knowledge. As only half of the students completed the first Hypothesis Testing Task (requiring a large one-sample z test), and two students the second (requiring a small two-sample t test), evidence of students' conceptual knowledge again depended very much on what they said, rather than what they did during the tasks.

FINDINGS AND DISCUSSION

Initial examination of the talk aloud and interview results showed that while some students had a reasonable conceptual knowledge of some concepts, many seemed to have a more limited understanding. A widespread difficulty was associated with the expression of statistical ideas. These initial findings were examined further, influenced by the studies of McKeown and Beck (1990) and Perkins and Simmons (1988), who each characterised novices' knowledge of different subject areas, believing that students' problems could not simply be attributed to a lack of knowledge. For example, novices often have conceptual gaps in their repertoires of knowledge, but these repertoires also contain much incorrect knowledge.

Table 1 summarises the main responses relating to conceptual knowledge of the four concepts. Predominantly, students' protocols were definitions or statements about relationships. Sometimes, definitions were expressed in terms of relationships. In Table 1, acknowledgement of an idea meant that a student offered a definition of the particular concept, and this was classified as correct, partially correct, or incorrect. A correct classification meant that the idea was conveyed with statistical accuracy, for example, "you could have as the null hypothesis the mean of the population being equal to 180... so the alternative hypothesis could be that the mean equals less than 180" (correct example). A partially correct classification usually included statistically imprecise information or

gaps in the student's knowledge about the concept, for example, "p-value is the probability of statistic correct if null hypothesis is correct" (partially correct definition). An incorrect classification meant that a student had the wrong idea, possibly a misconception, for example, "I would say the significance level is ultimately say a variance ... an ultimate line on the *x* and *y* axis" (incorrect definition). Acknowledgement of a relationship meant that a connection between a particular concept and another one provided on the Concept Mapping Task was noted by the student, for example, "p-value is a probability" (relationship between p-value and probability, no elaboration). When the connection was elaborated upon, and not simply acknowledged, this too was rated as correct, partially correct, or incorrect (with the same meanings as before). For example, in the following statement, "the statistic might be very significant if you have a high p-value," the significance and p-value concepts are linked incorrectly (relationship with incorrect elaboration).

Table 1
Summary of Number of Acknowledgements of the Major Ideas and Relationships, and their Classifications

Concepts		Number of Acknowledge-ments	Classification of Acknowledgements or Elaborations			
			Correct	Partially Correct	Incorr.	**Totals**
Hypothesis	Major Ideas	**71**	17	48	6	**71**
	Relationships	**76**	19	49	7	**75**
Significance level	Major Ideas	**34**	13	18	3	**34**
	Relationships	**32**	8	13	6	**27**
P-value	Major Ideas	**19**	0	3	5	**8**
	Relationships	**34**	11	7	6	**24**
Significance	Major Ideas/ Relationships	**16**	1	8	1	**10**
TOTALS		**282**	**69**	**146**	**34**	**249**
				249		

Table 1 exhibited three major features. First, the best-known concept was the hypothesis concept. Collectively, there were more statements made about the hypothesis concept (71+75) than any other concept in the table. This may account for the poor responses on the Hypothesis Testing tasks, many of which did not progress beyond the hypothesis statements. The table shows that the students found the other concepts considerably more difficult to discuss, with the total number of statements becoming fewer for concepts progressively lower in the table. The least known concept was the significance concept, and information contributed about this concept was mainly in the form of relationships. Several students could contribute no information at all about a concept (respectively 3, 3, and 7 students for significance level, p-value, and significance).

Second, there were more "Acknowledgements" classifications than "Correct," "Partially Correct," or "Incorrect" classifications (282 versus 249). This difference was evident mainly in the Relationships statements, but also in the Major Ideas associated with the p-value and significance concepts. The implication from this finding was that while students had some sense of the concepts and those with which they are related, on many occasions they had difficulty elaborating the nature of the relationships.

Third, "Partially Correct" or "Incorrect" statements were more common than "Correct" ones (146+34 versus 69). This implied that even when students could discuss an idea or elaborate upon a relationship, it was rarely statistically accurate.

In order to understand the nature of students' conceptual difficulties with the above concepts, further analysis was undertaken. When statements about a particular concept were not classified as "Correct," the student's responses associated with that concept were examined across all tasks. A combination of the categories used by McKeown and Beck (1990) and Perkins and Simmons (1988) was found to be useful in characterising students' faulty knowledge. It was found that in this study, faulty conceptual knowledge could be classified as *incomplete knowledge* (no elaboration, gaps in knowledge), *garbled knowledge* (knowledge about one concept is mixed up with knowledge about another concept), *more sophisticated knowledge* (but with some errors), and *misconceptions* (wrong ideas). Another difficulty, lack of precision in expressing statistical ideas, was usually classed as garbled knowledge, but sometimes it could also be associated with incomplete knowledge. That is, it was sometimes difficult to determine conclusively which category was most appropriate when students were statistically imprecise in their explanations. Table 2 summarises these classifications, provides a comment explaining each one, and gives the number of students with each type of problem for all four concepts.

Table 2

Conceptual Problems Associated with Hypothesis Testing Concepts and the Number of Students Exhibiting the Problem for Each Concept

Problems and comments	Concepts			
	H	SL	P	S
Incomplete knowledge	18	12	8	11
• no qualifications in using words such as "accept", "prove", "true", "opposite"	*			
• lack of precision in expressing statistical ideas	*			
• no elaboration	*	*	*	*
• gaps in knowledge	*	*	*	*
Garbled knowledge	8	13	10	5
• confused associations	*	*	*	*
• lack of precision in expressing statistical ideas		*	*	*
• confusing statistical language			*	
• problems with interpreting numbers		*	*	*
More sophisticated knowledge	6	7	3	0
• generally good conceptual knowledge, but with some lapses in the above (often expression)	*	*	*	
• visual representations			*	
Misconceptions	5	2	1	0
• incorrect ideas	*	*	*	

Note: H – hypothesis; SL – significance level; P – p-value; S – significance; * - difficulty evident for this concept

Incomplete knowledge was a major problem for the hypothesis and significance concepts, slightly less so for the significance level and p-value concepts. For the hypothesis concept, two concerns were that words such as "accept", "prove" and "true" were unqualified, and that statistical expression concerning the null and alternative hypotheses and the relationships described with other concepts were poor. Both are perhaps unavoidable in novices, because of the newness of the language and the abstract nature of the concepts. However, instruction should highlight the statistical inaccuracies associated with the first concern in particular, because it is part of the knowledge pertaining to the hypothesis concept. With respect to the second concern, imprecise statistical expression detracted

from the ideas being presented, and usually consisted of partially correct statements. For the hypothesis concept, this usually meant that relationships were not initially explained correctly, nor were they enhanced through additional explanations or actions at another time. This meant that there were gaps in the knowledge base. For all concepts, lack of elaboration of ideas, and gaps in knowledge, were problems associated with incomplete knowledge. This usually meant that the particular concept was underdeveloped, or there was no evidence of any knowledge about it at all. Furthermore, gaps in knowledge usually meant that the concept was not included on the students' concept map.

Garbled knowledge was the main problem associated with the significance level and p-value concepts, less so for the hypothesis and significance concepts. One difficulty associated with garbled knowledge was confused associations between concepts. Examples include failure to differentiate between the separate roles of the p-value, z value, and t value, and mixing the meaning of significance level with significance, p-value, critical region, and confidence interval. This may be a result of the number of concepts associated with hypothesis testing, whose definitions and roles become jumbled together because of the pace of the subject and the large amount of content in statistical subjects. Care must be taken by instructors that text books do not cause this confusion through inadequate definitions. Again, problems with the language of statistics may be avoided by creating an environment where statistical ideas and concepts are openly discussed.

Another difficulty associated with garbled knowledge was imprecise statistical expression. When this was included in garbled knowledge, it usually meant that relationships were not initially explained correctly, yet at some later stage in the interview, a student showed understanding of the relationships. This normally occurred through performance on the Hypothesis Testing Tasks or through demonstration on a distribution diagram, a visual representation. As the Hypothesis Testing Tasks presented a major problem to many students, this avenue for explaining or highlighting relationships was closed to them. Hong and O'Neil's (1992) study on statistics students concluded that emphasising distribution diagrams in instruction facilitated students' understanding of the concepts. The present study provides some support for this conclusion, particularly in its facility for clarifying relationships between concepts when language is inadequate.

Confusing statistical language was another difficulty included in the classification of garbled knowledge, and this was concerned with the different meanings associated with extreme ideas such as high p-values, high significance, and high significance level. Visual representation in the form of distribution diagrams would assist in establishing the differences between the three concepts of p-value, significance, and significance level. Creating an environment for discussing concepts could help reduce such incorrect usage.

Garbled knowledge also included evidence of difficulties associated with interpreting the numerical values, and this applied mainly to the p-value concept. One example suffices as an illustration: "if it's very close to zero then it's good, but if it's getting, usually over the 20 mark, above that I think is kind of shady on whether to accept it or not." Undoubtedly, this difficulty is linked to other difficulties such as confused associations with other concepts.

More Sophisticated Knowledge was present for the hypothesis, significance level, and, to a lesser extent, the p-value concepts. It meant that students demonstrated a better grasp of the concept than other students, particularly in terms of the elaboration of the relationships with other concepts. Nevertheless, there were still occasional lapses in statistical precision and incidences of garbled or incomplete knowledge. For example, for the p-value concept, students had difficulty explaining the link with probability, but students in this category represented p-value diagramatically, aptly demonstrating and explaining the numerical values associated with p-value.

Misconceptions were the least common problem. It is arguable whether some of these really were misconceptions, or just erroneous attempts to say something to the researcher at the time. For the significance level concept, one misconception was associated with its definition as the probability of being wrong, and another as being the same as Type I error. These misinterpretations could be a result of confusion of the probabilities associated with the concept (Falk, 1986). For the p-value concept, the single misconception was associated with the contention that p-value is always low. This misinterpretation existed because the examples this particular student actually performed, or was shown as models, resulted in low p-values. Such a misconception can easily be eradicated through carefully chosen problems which result in a mixture of low and high p-values, and through the discussion of such results.

Given the lack of consensus on definitions, terminology, and interpretations cited in the statistical literature, it is hardly surprising that many problems of this nature exist in novice students after a one-semester course of statistical study. In response to why students do not seek out ways of improving their knowledge, Perkins and Simmons (1988) suggested that students' inquiry frame is not cultivated. First, they are often not encouraged to seek out and ponder upon problems that might improve their knowledge. Second, the applications of the topic to be learned are situated solely in an academic environment where students master text book abstractions. Third, there is little encouragement to develop knowledge beyond the boundaries of the content area with "What if?" and "What if not?" questions.

CONCLUSIONS

The results in this study revealed that students have a major problem in expressing statistical ideas with accuracy (see Table 1). Sometimes this masks their otherwise good conceptual knowledge of hypothesis testing concepts, and at other times it is representative of their inadequacies. A solution to this problem by lecturers may mean facilitating situations (seminars, reflections, discussions, explorations, critical evaluations, writing tasks, concept mapping) where students have no option but to conquer their fears about statistics and immerse themselves in the utilisation of statistical language. This would encourage the organisation of ideas, and improve the clarity of explanations. As Knapp (1996) noted, one of the annoying things about hypothesis testing is the jargon associated with it, and the sooner it is taken on board the better.

Furthermore, concept mapping proved to be useful for investigating conceptual knowledge. It was effective in studying how people link and structure ideas and relationships relating to concepts, in highlighting the main problems, and in establishing whether key concepts in a topic were known. These functions make the concept map a powerful tool. In particular, it can invoke discussion about concepts and their relationships, especially when combined with talking aloud and interview methods. However, as indicated above, the concept map used on its own may not be as useful, because students have difficulty adding the linking words on the map itself. Its true potential is realised when combined with other methods. Then it also becomes a powerful learning tool.

REFERENCES

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education*, *61*(4), 287-292.

Chandler, R. E. (1970). The statistical concepts of confidence and significance. In D. E. Morrison & R. E. Henkel (Eds.), *The significance test controversy* (pp. 213-215). Chicago: Aldine. (Original work published 1957 in *Psychological Bulletin*, *54*(5), 429-430).

Chi, M. T. H., & Koeske, R. D. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology*, *19*(1), 29-39.

Ehrenberg, A. S. C. (1990). A hope for the future of statistics: MSOD. *The American Statistician, 44*(3), 195-196.
Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning, 9*, 93-96.
Freund, J. E., & Perles, B. M. (1993). Observations on the definition of p-values. *Teaching Statistics, 15*(1), 8-9.
Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*(4), 379-390.
Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19*(1), 44-63.
Gigerenzer, G. (1993). The superego, the ego, and the id in statistical thinking. In G. Keren & C. Lewis (Eds.), *Handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum.
Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*(1), 15-24.
Hawkins, A. (1991). Success and failure in statistical education - a UK perspective. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (Vol. 1, pp. 24-32). Voorburg, The Netherlands: International Statistical Institute.
Hawkins, A., Jolliffe, F., & Glickman, L. (1992). *Teaching statistical concepts*. London: Longman.
Hong, E., & O'Neil, Jr., H. (1992). Instructional strategies to help learners build relevant mental models in inferential statistics. *Journal of Educational Psychology, 84*(2), 150-159.
Huberty, C. J. (1985, April). *On statistical testing*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
Knapp, T. R. (1996). *Learning statistics through playing cards*. London: Sage.
McKeown, M. G., & Beck, I. L. (1990). The assessment and characterization of young learners' knowledge of a topic in history. *American Educational Research Journal, 27*(4), 6988-726.
Menon, R. (1993). Statistical significance testing should be discontinued in mathematics education research. *Mathematics Education Research Journal, 5*(1), 4-18.
Perkins, D. N., & Simmons, R. (1988). Patterns of misunderstanding: An integrative model for science, math, and programming. *Review of Educational Research, 58*(3), 303-326.
Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *The Journal of Experimental Education, 61*(4), 350-360.
Thompson, B. (1994). *The concept of statistical significance testing*. (ERIC Document Reproduction Service No. ED 366 654)
Truran, J. M. (1998). The development of the idea of the null hypothesis in research and teaching. In L. Periera-Mendoza, L. S. Kea, T.W. Kee, & W. Wong (Eds.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 1067-1073). Voorburg, The Netherlands: ISI Permanent Office.
Watts, D. G. (1991). Why is introductory statistics difficult to learn? And what can we do to make it easier? *The American Statistician, 45*(4), 290-291.
West, L. J. (1990). Distinguishing between statistical and practical significance. *Delta Pi Epsilon Journal, 32*(1), 1-4.