ESTABLISHING PCK FOR TEACHING STATISTICS

Jane M. Watson[1], Rosemary A. Callingham[2] and Julie M. Donne[1]
[1]University of Tasmania, [2]University of New England, Australia
Jane.Watson@utas.edu.au

This report considers the pedagogical content knowledge (PCK) of 42 teachers selected to be part of a professional learning program in statistics. As part of a profile measuring many aspects of teacher confidence, beliefs, teaching practice, assessment practice, and background, PCK is addressed through responses to student survey items and how the items could be used in the classroom. Rasch analysis is used to obtain a measure of teacher ability in relation to PCK. Based on measured ability, three hierarchical clusters of teacher ability are identified, and the characteristics of each described in terms of the items likely to be achieved. These are exemplified with kidmaps of individual teachers' performances from each of the three clusters.

BACKGROUND

The official arrival of probability and statistics in the school mathematics curriculum around 1990 (e.g., National Council of Teachers of Mathematics, 1989) created a need for teachers to acquire the content knowledge and teaching practices required to deliver the topics. The early work of Hawkins, Jolliffe and Glickman (1992) was instrumental in providing background and a structure for considering the issues in teaching statistics. Much of their book was devoted to content, but there was no mention of research into teachers' understanding per se in relation to teaching statistics at the school level as it had not yet become an issue. In recognition of the emerging statistics curriculum in schools, the International Statistical Institute Roundtable of 1992 focussed on "Introducing Data Analysis in the School: Who Should Teach It and How?" (Pereira-Mendoza, 1993). At the time, Begg (1993) set a research agenda in which he included teacher professional development and questions about its effectiveness that implicitly, but not explicitly, involved teachers' knowledge for teaching statistics.

Over the 1990s there was an increasing interest in teachers' understanding of topics included in teaching statistics, how to teach statistics, and the professional development associated with teaching statistics. At the same time there was a growing appreciation in the field of teacher education of the various aspects of teacher knowledge related to successful teaching. This appreciation can be traced back to the seminal work of Shulman (1987) who listed seven types of teacher knowledge affecting teacher performance: (i) content knowledge; (ii) general pedagogical knowledge; (iii) curriculum knowledge; (iv) pedagogical content knowledge (PCK); (v) knowledge of learners and their characteristics; (vi) knowledge of education contexts; and (vii) knowledge of education ends, purposes, and values (p. 8). These types of knowledge were the basis of a profiling instrument devised by Watson (2001) for gauging the professional development needs of teachers and assessing the success of professional learning programs. This profile in turn provided a framework for the larger instrument within which the items discussed in this paper are embedded.

In recent times, the phrase PCK has come to be interpreted in a broader sense than perhaps first intended by Shulman. The increased breadth is related to the close connection recognised between content knowledge and PCK and between PCK and knowledge of learners and their characteristics. Hill, Rowan, and Ball (2005), appeared to incorporate these three types of knowledge to provide a rich description of "teachers' knowledge for teaching mathematics." Chick (e.g., 2007) has also extended the boundaries of PCK using a framework that includes links to content knowledge and students as learners, as well as curriculum knowledge, the purpose of content knowledge, and the goals for learning. This approach appears to treat PCK as the underlying and encompassing phrase to summarize Shulman's original intentions. The expansion is also reflected in the current study. Several of Shulman's types of knowledge are addressed in other parts of the larger survey completed by teachers as well as in the items presented here. The major focus of PCK in items in this study is teachers' content knowledge, its reflection in knowledge of their students' content knowledge, and their PCK in using student responses to devise teaching intervention. A similar approach was used by Watson, Beswick,

 In C. Batanero, G. Burrill, C. Reading & A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference.*

and Brown (2006) to measure teachers' PCK about fractions.

This study is part of the StatSmart project (Australian Research Council, Grant No. LP0669106), a professional learning program assisting teachers to appreciate the developmental processes that students go through in reaching statistical understanding, providing teachers with resources including software, and helping them devise learning activities suitable for their class levels. The project will last for three years and includes repeated monitoring of teachers and their students. Design details are discussed in Callingham and Watson (2007).

METHODOLOGY

Sample

Altogether, 42 teachers in the StatSmart project, located in three different states of Australia and in government, Catholic, and independent schools, completed the profile, 22 males and 20 females. Of the respondents, nearly half (n = 20) taught in high schools, including Grades 11 and 12. Of the others, eight taught only in primary grades, five were in a middle school setting (Grades 5 to 9), and nine taught in junior secondary schools (Grades 7 or 8 to 10).

Instrument

Teachers were administered a Teacher Profile instrument, based on one used earlier (Watson, 2001), consisting of six sections that addressed different aspects of Shulman's (1987) teacher knowledge. Background information about mathematics background, teaching experience, and grades taught was also collected. For this report, only the 12 items mainly addressing PCK are considered. These PCK items were based on student survey items used in earlier studies (e.g., Watson & Callingham, 2003) so that data were available about students' actual responses. The first set of eight items, based on three actual media articles in the Australian context, asked teachers to predict a range of responses their students might produce if presented with a question, and then to explain how they might use the question in their classrooms, including how they might intervene to address inappropriate responses. The topics addressed were odds (odda, oddb), based on a media headline about a sporting event (Figure 1); a piechart that summed to 128.5% (piea, pieb), based on a media article about Australian retailers; and graphing an association (roba, robb) and questioning a claim (robc, robd) based on an article that claimed an almost perfect relationship between motoring and heart deaths.

The second set of four items asked teachers to respond to actual student responses to two questions. In both questions two responses showed different levels of understanding of the core topic. The first was in the context of a classroom question about drawing a coloured ball from one of two boxes with different numbers but the same proportion of two different colours (boxa, boxb), and the second used a two-way table about lung cancer and smoking where no association was shown (taba, tabb). Teachers' responses were coded using a hierarchical scoring rubric that took account of the appropriateness of the response in terms of both mathematical understanding and contextual information (Figure 1). Student responses to the items indicated a full range of responses was obtained at all grade levels, so teachers in primary classrooms could be expected to recognise high level responses and were not disadvantaged.

Analysis

The data were analysed using Masters' (1982) Partial Credit Model (PCM) with Quest computer software (Adams & Khoo, 1996). The PCM is an extension of the Rasch model (Rasch, 1980) in which the interactions between persons and items are placed on the same measurement scale. The scale so produced is a genuine interval scale that allows comparison of person performance on the set of items used (Bond & Fox, 2007). This model is appropriate for the kind of data obtained here because the hierarchical nature of the scoring rubric provided direction along a single variable of PCK. The distance between each step of the coding is not assumed to be identical, and different numbers of steps in each item can be accommodated.

The quality of the instrument used is assessed using the fit to the model. If all items provide good fit to the PCM, then the instrument is assumed to be measuring a single construct, in this instance PCK. The usual measure of fit reported is the infit mean square statistic, which has an ideal value of 1.00. Acceptable levels of fit lie between 0.77 and 1.3 (Keeves &

Alagumalai, 1999). The standardised fit measure provides a *z*-statistic, providing the statistical significance of the fit figure, using the usual accepted values of ± 1.97. The Person Separation Reliability indicates the extent to which the set of items separates the persons along the scale. It has an ideal value of 1, and values above approximately 0.7 provide acceptable separation, allowing persons to be compared on the basis of their measured ability. Estimates of person ability were obtained in logits (the logarithm of the odds of success). Teachers were rank ordered on the basis of the ability estimates obtained. Using the differences in ability as a basis, teachers were divided into three groups, labelled low, middle, and high in terms of PCK.

The Quest computer program produced a "kidmap" for each teacher. Kidmaps show all items attempted by an individual in terms of items correct and incorrect, related to the estimated ability of the person. Figure 2 shows an example of a kidmap. The XXX shows the estimated ability of the person, measured at -1.68 logits, which places this teacher in the low grouping. The dotted horizontal lines indicate the measurement error associated with this estimate. Items that occur within the range of measurement error are those on which the person has about a 50% chance of succeeding. Items to the right of the map are those that have not been successfully answered. Those on the left are items on which this teacher has been successful. As the items become harder, the teacher has a decreasing probability of success. Items appearing in the bottom right hand quadrant are easy items that are unexpectedly wrong. In the map shown, there are none of these items. In the upper left hand quadrant, harder items appear that are unexpectedly correct. There are two items in this quadrant. This teacher is behaving acceptably in relation to the model, shown by the fit value of 0.96, which lies within the accepted range of 0.77 to 1.3. Fit values below 0.77 indicate highly predictable patterns of response.

The following question came from Student Surveys used in research:

North at 7-2
But we can still win match, says coach

What does "7-2" mean in this headline about the North against South football match? Give as much detail as you can.
From the numbers, who would you expect to win the game?

Odda. What kinds of responses would you expect from your students? Write down some appropriate and inappropriate responses.
0 = No response
1 = Response not addressing odds or the context of the question
2 = Response indicating either a correct approach *or* an incorrect approach
3 = Response containing both appropriate and inappropriate approaches to the problem

Oddb . How would/could you use this item in the classroom? For example, how would you intervene to address the inappropriate responses?
0 = No response
1 = Response not addressing the mathematical content of the problem
2 = A single generic idea or ideas for the problem
3 = Reference to two or more ideas without linking them
4 = Discussion including reference to odds and correct interpretation with specific examples

Figure 1. Question and scoring rubric (odda, oddb)

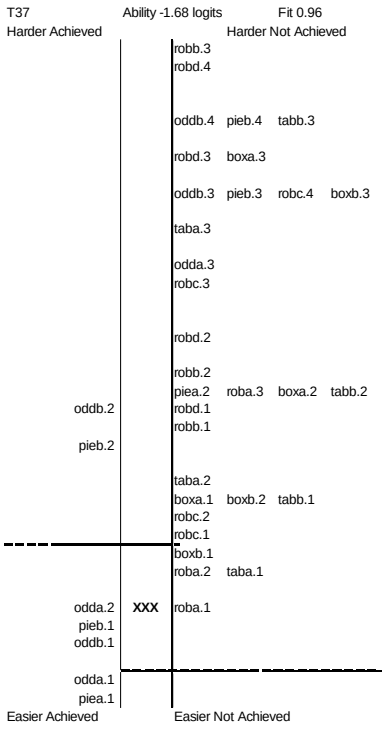

Figure 2. An example of a low ability map

RESULTS

Overall fit to the model of all items was good ($IMSQ_I$ = 1.01, *z* infit = 0.08). In addition, all individual items showed fit to the model, indicating that the instrument provided a defensible measure of a single construct, labelled PCK. The Person Separation Reliability figure was 0.77, indicating acceptable separation of people along the scale. The overall person fit was also good ($IMSQ_P$ = 1.04, *z* infit = 0.05). The mean estimated ability of all persons was 0.13 logits compared with the item mean difficulty of 0.00, suggesting that the items were reasonably well matched to the teacher group. These overall statistics indicated that the scale provided a reliable

basis on which to base teacher comparisons.

On the basis of the individual measured abilities, the teachers were divided into three groups. The low group included 14 teachers, the middle group, 19 teachers, and the high group, nine teachers. A one-way ANOVA indicated that the mean differences in measured ability between the groups were significant at the 0.01 level (df 2, $F = 57.388$, $p = 0.00$), suggesting that these groupings were justifiably different. The kidmaps for all teachers were obtained and sorted into the three groups. Each group was then examined for similarities and differences to identify commonalities and exceptions within each grouping, and differences between groups.

"Low" Level

In general, teachers in the low grouping were likely only to be partially successful on items that requested them to suggest students' responses and then to indicate how they would address these responses in the classroom. They were unlikely to demonstrate any success on items that asked them to respond to actual student answers to questions addressing proportional reasoning. This finding suggests that these teachers could only begin to predict students' responses and to use materials in the classroom. Two examples are presented.

Teacher T37 had the lowest ability estimate of -1.68 (see Figure 2). The fit (0.96) was acceptable. The only items on which the teacher achieved a non-zero code were related to the pie chart and odds. For the odds question only incorrect student responses were suggested for students (odda.2): "Game response. Goals only method of scoring... The state of the game is on a break or interval?" When addressing how the question could be used in the classroom T37 said (oddb.2): "Highlight key language. Interpret data only... Display data by alternate means..." This response was at a higher level than expected given T37's overall performance. All other questions, including four in the region with approximately a 50% chance of success were not achieved at the lowest level coded.

T16 had an ability estimate of -0.61 and a fit value of 0.44, indicating overfit. Two harder items achieved were close to the 50%-chance region, and no easier items were achieved. Both achieved items were related to expected student responses. Figure 3 shows expected correct (left) and incorrect (right) graphs of the nearly perfect relationship of motoring and heart deaths (roba.3). T16 gave examples of both appropriate and inappropriate student responses for the pie chart question (piea.2). Both the error in the graph and other graphical observations or context questions were given: "The Other grocery stores together have more shares than any of Coles, Davids, IHL & Woolworths... The percentages don't add up to 100%."

Figure 3. T16 response

"Middle" Level

Teachers in the middle grouping were likely to be able to suggest both correct and incorrect responses for the graphs of the motoring/heart deaths claim and to find the error and make other suggestions for the pie chart item. For using responses to develop ideas for the classroom, however, they were likely to suggest single generic ideas for the two scenarios. Most of these teachers could respond to the student answers to the proportional reasoning problems with questions involving mathematical content.

The teacher chosen to demonstrate an acceptable fit (0.96) in the middle level was Teacher T32 with an ability estimate of 0.60, the same as three other teachers at the top of the middle group. This teacher behaved somewhat differently from those discussed thus far in struggling with the motoring/heart death items except for suggesting students' questions about the claims. Here T32 made the following suggestions (robc.4): "Does one cause the other? Which?... What other factors show the same pattern? What is the nature of the connection between these two factors?" Except for two items, all other responses achieved Code 2. The two easier items not achieved were related to the motoring/heart death question.

Another teacher in the middle level, T08, had an ability estimate of 0.04 but with a fit value (0.35) indicating overfit in the responses (see Figure 4). This is seen in only one harder item unexpectedly achieved, suggesting student questions about claims in the motoring/heart

deaths scenario (robc.3): “Who did you survey? How large was the group? Did you survey particular age groups? Which ones?...” Of the 10 items in the central region with a 50% chance of success, T08 was successful on four.

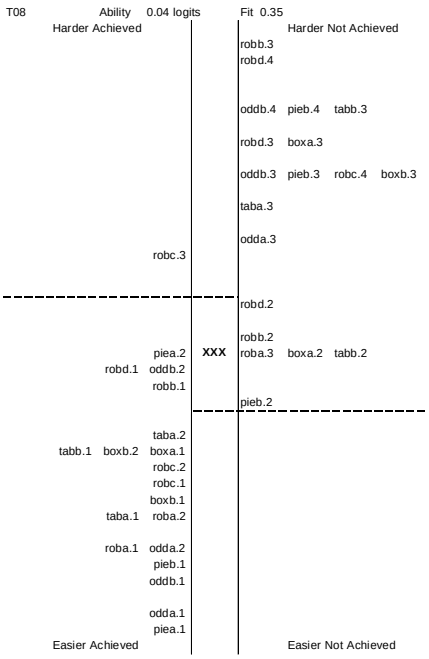

Figure 4. An example of a middle ability kidmap

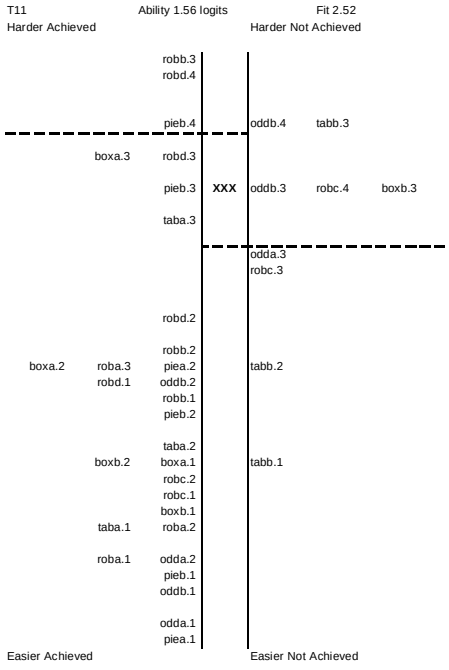

Figure 5. An example of a high ability kidmap

“High” Level

In the high grouping many teachers still had difficulty achieving the highest codes on individual items, especially those related to using students’ responses as a basis for planning intervention activities. Overall, however, they showed a relatively high likelihood to focus on the mathematics involved in the proportional reasoning problems and suggested both correct and incorrect responses to the media items. The second highest ability estimate in the high group was for Teacher T11, with an estimate of 1.56. The fit value (2.52) indicated random behaviour seen in Figure 5, with four easier items not achieved and three harder items achieved. T11 reached the highest code for both parts of the pie chart item (piea.2 and pieb.4), the responses to the student answers to the first proportional chance item (boxa.3) and the first conditional table item (taba.3). For the proportional chance item (boxa.3), T11 responded with the questions, “Are you more likely to get a blue from one of the boxes? If you draw 100 times from each box what would you expect?” and for the conditional table (taba.3), the response was, “Is 90/150 the same as 60/100? Could model the question using counters... Do trials.”

The highest achieving teacher, T14, had an ability estimate of 2.34, with a fit value (0.65) indicating overfit. Only three items were not achieved at the highest level, with one (robd) unexpectedly not achieved at Code 3. Except for parts of items related to the motoring/heart deaths article and the use of the pie chart question in the classroom, all items were achieved at the highest code. For example, for the second proportional chance item (boxb.3), T14 responded, “Are there also more blues in B? If I have less marbles do I have a better chance always?” and for the second conditional table item (tabb.3), “... how could the chance of lung disease be best expressed from the sample. Can we use a percentage?”

DISCUSSION

This study presents an initial attempt to characterise teachers’ PCK with respect to understanding required to teach statistics. The study considered three groups of teachers, identified as low, middle and high according to their measured ability against a set of items that addressed PCK. It was noticeable that teachers who unexpectedly did not achieve easy items often missed items requiring a response to a specific student misunderstanding. It seems that

assuming teachers can identify the next steps to move students towards higher levels of statistical understanding may be misplaced, even in teachers with high levels of PCK. This has implications for professional development programs, which may need to focus more clearly on developing targeted intervention with respect to students' levels of understanding. Teachers unexpectedly achieving harder items had more mixed responses; all types of items appeared in this category, which may reflect the idiosyncratic nature of individual teachers' experiences.

This study represents an initial attempt to establish the nature of teachers' demonstrated PCK. Over the period of the StatSmart project, teachers will complete a similar survey three times so that changes in their statistical understanding can be monitored, and related changes in their PCK identified. StatSmart is funded by the Australian Research Council, LP0669106.

REFERENCES

Adams, R. J., & Khoo, S. T. (1996). *Quest: Interactive item analysis system. Version 2.1* [Computer software]. Melbourne: Australian Council for Educational Research.

Begg, A. (1993). Establishing a research agenda for statistics education. In L. Pereira-Mendoza (Ed.), *Introducing data analysis in the schools: Who should teach it and how?* (pp. 212-218). Voorburg, The Netherlands: International Statistical Institute.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2 ed.). Mahwah, NJ: Lawrence Erlbaum.

Callingham, R., & Watson, J. (2007). Overcoming research design issues using Rasch measurement: The StatSmart project. In P. Jeffery (Ed.) *Proceedings of the Australian Association for Research in Education annual conference, Fremantle, 2007*. Online: www.aare.edu.au/07pap/cal07042.

Chick, H. L. (2007). Teaching and learning by example. In J. Watson & K. Beswick (Eds.), *Mathematics: Essential research, essential practice: Proceedings of the 30th annual conference of the Mathematics Education Research Group of Australasia, Hobart* (pp. 3-21). Adelaide, SA: MERGA.

Hawkins, A., Jolliffe, F., & Glickman, L. (1992). *Teaching statistical concepts.* New York: Longman Publishing.

Hill, H. C., Rowan, R., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Keeves, J. P., & Alagumalai, S. (1999). New approaches to measurement. In G. N. Masters and J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 23-42). Oxford: Pergamon.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

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

Pereira-Mendoza, L. (Ed). (1993). *Introducing data analysis in the schools: Who should teach it and how? Proceedings of the International Statistical Institute Round Table Conference*. Voorburg, The Netherlands: International Statistical Institute.

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

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1-22.

Watson, J. M. (2001). Profiling teachers' competence and confidence to teach particular mathematics topics: The case of chance and data. *Journal of Mathematics Teacher Education, 4*, 305-337.

Watson, J., Beswick, K., & Brown, N. (2006). Teachers' knowledge of their students as learners and how to intervene. In P. Grootenboer, R. Zevenbergen, & M. Chinnappan (Eds.), *Identities, cultures and learning spaces*: *Proceedings of the 29th annual conference of the Mathematical Education Research Group of Australasia, Canberra* (pp. 551-558). Adelaide, SA: MERGA.

Watson, J. M., & Callingham, R. A. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3-46. Online: http://www.stat.auckland.ac.nz/~iase/serj