

Explaining Student Achievement: the Influence of Teachers' Pedagogical Content Knowledge in Statistics

Rosemary Callingham¹ · Colin Carmichael² ·
Jane M. Watson³

Received: 17 June 2014 / Accepted: 2 May 2015 / Published online: 3 June 2015
© Ministry of Science and Technology, Taiwan 2015

Abstract Statistics is an increasingly important component of the mathematics curriculum. *StatSmart* was a project intended to influence middle-years students' learning outcomes in statistics through the provision of appropriate professional learning opportunities and technology to teachers. Participating students in grade 5/6 to grade 9 undertook three tests, a pre-test, a post-test and a longitudinal retention test over a period of 2 years. Their teachers completed a survey that included items measuring pedagogical content knowledge (PCK) for teaching statistics. Despite the development of valid instruments to measure both student and teacher content knowledge and teachers' PCK, linking teachers' knowledge directly to students' learning outcomes has proved elusive. Multilevel modelling of results from 789 students for whom there were 3 completed tests and measures from their teachers indicated that students' outcomes were influenced positively by their initial teacher's PCK. Extended participation of teachers in the project also appeared to reduce negative effects of changing teachers.

Keywords Middle years · Pedagogical content knowledge · Statistics · Student achievement

✉ Rosemary Callingham
Rosemary.Callingham@utas.edu.au

Colin Carmichael
ccarmichael@csu.edu.au

Jane M. Watson
Jane.Watson@utas.edu.au

¹ School of Education, University of Tasmania, Locked Bag 1307, Launceston, TAS 7250, Australia

² School of Education, Charles Sturt University, PO Box 789, Albury, NSW 2640, Australia

³ School of Education, University of Tasmania, Private Bag 66, Hobart, TAS 7001, Australia

Introduction

Since the last decade of the twentieth century, statistics has had a defined place within the school mathematics curricula of many countries, including the USA (National Council of Teachers of Mathematics, 1989), Australia (Australian Education Council (AEC), 1991), and New Zealand (Ministry of Education [MENZ], 1992). Although there have been some revisions to the suggested content over the years (e.g. Australian Curriculum, Assessment and Reporting Authority, 2011; Franklin et al., 2007; MENZ, 2007; National Council of Teachers of Mathematics, 2000), the place of statistics has been consolidated and a research field of statistics education established. The progress of this field has been summarised in the reviews of Konold & Higgins (2003) and Shaughnessy (2007). As progress was made on documenting students' developing understanding, the scope widened to include teachers' understanding of statistics and their needs to facilitate successful implementation of the statistics component of the mathematics curriculum in the classroom. The many issues surrounding the complexities of translating the curriculum to successful learning outcomes for students via teachers in classrooms were explored in detail by Batanero, Burrill & Reading (2011) following an intensive International Commission on Mathematical Instruction (ICMI) and International Association for Statistics Education (IASE) joint study in 2008. Acknowledging that many teachers had not studied statistics in their own education and training, many of the studies looked at the pre-service and in-service learning of teachers and the measurement of their improvement in understanding content and pedagogy (e.g. Callingham & Watson, 2011; Makar & Fielding-Wells, 2011; Pierce & Chick, 2011).

Of concern in this paper is the complicated relationship between teachers' pedagogical content knowledge (PCK) for teaching statistics and students' learning outcomes as they develop understanding of statistical concepts. As part of a complex 3-year study that took place between 2007 and 2010 associated with a professional learning programme for teachers in statistics, instruments were developed and/or adapted to measure various aspects of teachers' knowledge for teaching statistics and of students' understanding of the concepts (Callingham & Watson, 2007). This paper reports on teachers' initial levels of PCK for teaching statistics and the change in student understanding as measured at three points in time: before exposure to a unit on statistics, after the unit, and 1 year later.

Measuring Teachers' Knowledge

Interest in measuring teachers' knowledge for teaching mathematics has developed since Shulman's (1987) general categorization of seven types of teachers' knowledge required for successful teaching: curriculum knowledge, general pedagogical knowledge, content knowledge, pedagogical content knowledge, knowledge of learners and their characteristics, knowledge of education contexts, and knowledge of education ends, purposes, and values. Since Shulman published his seminal work, many approaches have been taken, and attempts have been made to describe, characterise, and measure teachers' mathematics knowledge. Ma (1999) described elementary teachers' deep connected mathematics knowledge as Profound Understanding of Fundamental

Mathematics (PUFM). Even & Tirosh (2002) went beyond mathematics knowledge to consider teachers' knowledge of students' mathematical learning.

Other researchers have worked with or adapted Shulman's framework for the teaching of mathematics or statistics. For example, Watson (2001) developed a profile instrument to measure teachers' knowledge on each of Shulman's dimensions in relation to the teaching of data and chance. This instrument was later broadened for the teaching of middle school mathematics more generally (Beswick, Callingham & Watson, 2012; Watson, Brown, Beswick & Wright, 2011).

A major contribution to the field was that of Hill, Ball and colleagues (e.g. Ball, Thames & Phelps, 2008; Hill, Schilling & Ball, 2004; Hill, Sleep, Lewis & Ball, 2007). They adapted Shulman's work, originally focussing on the mathematics knowledge required but later expanding their work to acknowledge other necessary aspects: common content knowledge, specialised content knowledge, knowledge at the mathematical horizon, knowledge of content and students, knowledge of content and teaching and knowledge of curriculum. These components encompass what others have continued to recognise as pedagogical content knowledge including an implicit appreciation of students as learners (Callingham & Watson, 2011) and the recognition of the particular affordances of tasks chosen by teachers for use in the classroom (Chick, 2007). More specifically in relation to teaching statistics, Groth (2007) synthesised the work of Ball and her colleagues into four categories for the statistics classroom: common knowledge, specialised knowledge, mathematical knowledge and non-mathematical knowledge, implicitly recognising the importance of context in statistics (Callingham, Watson & Burgess, 2012).

In the *StatSmart* project, pedagogical content knowledge was framed within Shulman's (1987) original definition:

the blending of content and pedagogy into an understanding of how topics, problems, or issues are organized, represented, and adapted to the diverse interests and abilities of learners, and presented for instruction. Pedagogical content knowledge is the category most likely to distinguish the understanding of the content specialist from that of the pedagogue. (1987, p. 8)

This definition was chosen because it was closely aligned to the aims of the study, which were focussed on the overall improvement of statistics teaching rather than attempting to develop a fine-grained description of the nature of teachers' knowledge for teaching statistics. Although it was not the original intention behind Shulman's work, with the increasing pressure for accountability in schooling and the scale of the project, there was also interest in providing solid quantitative data about both teachers' knowledge and students' learning outcomes in statistics.

Within Shulman's broad definition of PCK, given teachers' lack of depth in statistical understanding (Batanero, Burrill & Reading, 2011), recognising students' misconceptions and strategies to remediate these became a focus of *StatSmart*, with the intention of improving the quality of statistics teaching in the project schools. Instruments were developed to measure teachers' statistical PCK that considered (i) their prediction of students' likely answers to statistical problems; (ii) teachers' responses to students' actual answers taken from student surveys; and (iii) teachers' intervention strategies in relation to students' current knowledge (see Callingham & Watson, 2011 for further

details). As such, these instruments addressed many of the aspects of Ball et al.'s (2008) conceptualisation of teachers' knowledge, and Groth's (2007) framework for considering teachers' statistical knowledge. For example, items relating to prediction of correct and incorrect responses drew on both teachers' own statistical understanding, without which they could not predict high-level responses, and their specialised knowledge of statistics in the classroom, which they needed to identify students' common misconceptions. Time constraints on teachers prevented the administration of more nuanced instruments because these would have required a longer survey with more items to address clearly the different domains. Hence, a "thick" construct of teachers' statistical PCK was the target variable used in the *StatSmart* study.

Measuring Student Understanding

In contrast to the measurement of teacher knowledge, measuring student knowledge has a longer history. The measurement of student understanding of statistical concepts dates back to ideas associated with "average", usually interpreted as the arithmetic mean, as well as concepts in probability. Both of these topics were typically found in earlier curriculum documents associated with procedures for calculating means and probabilities. In the 1980s, Pollatsek, Lima & Well (1981) and Mevarech (1983) demonstrated student difficulty with weighted averages, as did Strauss & Bichler (1988) with the general properties of the mean. Mokros & Russell (1995) identified the dilemma of representativeness for averages, and Cai (1995, 1998) revealed difficulties with the notion of mean due to students' failure to work the algorithm backward. Similarly, Green (1983, 1986, 1991) produced the first large-scale longitudinal research in the related area of probability.

With the advent of the statistics component of national curricula, interest in measuring student understanding over a broader range of statistical ideas grew, for example including sampling (Watson & Moritz, 2000) and beginning inference (Watson & Moritz, 1999). The work of Watson and colleagues was consolidated in a scale of statistical literacy (Watson & Callingham, 2003) based on student surveys, and in a scale of statistical understanding reflecting adoption of the concepts of variation and expectation (Watson, Callingham & Kelly, 2007) based on in-depth student interviews. Many of the items used in these studies drew on the earlier work of other researchers, such as Batanero and her colleagues (e.g. Batanero Estepa, Godino & Green, 1996).

The *StatSmart* study drew on this body of work to develop instruments to measure students' statistical understanding. Items included many that had been used in prior studies (e.g. Watson & Callingham, 2003) together with a small number of new items to expand the item pool and provide additional information about specific statistical ideas.

Relationship Between Teacher Knowledge and Students' Outcomes

Despite the activity on developing instruments and identifying key aspects of teacher knowledge, it has been surprisingly difficult to link teacher knowledge directly to students' learning outcomes. It has long been recognised that using proxy measures of teachers' mathematical knowledge, such as qualifications or training experience, shows no relationship to students' learning outcomes in mathematics (e.g. Mewborn, 2001).

A major contribution to the field was made by Hill and colleagues (Hill et al., 2004; Hill, Rowan & Ball, 2005) who unpacked ideas about teachers' specialised mathematical knowledge and developed an instrument to measure elementary teachers' mathematical knowledge for teaching. This measure included actions such as providing examples, explaining concepts, correcting work and using a range of representations of mathematical ideas. They found that teachers' "knowledge of mathematics for teaching" predicted gain scores in two lower elementary grades.

More recently, a German study of a representative sample of grade 10 classrooms over 1 year identified that teachers' mathematics pedagogical content knowledge had a large positive impact on their own students' learning gains (Baumert et al., 2010). They identified that 39 % of the variance between classrooms was due to the variable they identified as PCK. Further, they indicated that the relationship was linear and that PCK was more important than content knowledge.

These findings suggest that classroom teachers understand mathematical ideas in specialised ways, and that this specialised knowledge has a positive impact on students' learning gains. In the study reported here, the context was statistics, rather than pure mathematics, and students' learning trajectories were considered using at least three data points. Rowan, Correnti & Miller (2002) have suggested that this approach avoids some of the difficulties associated with using learning gains. In addition, similar to the Hill et al. (2005) and the Baumert et al. (2010) studies, a direct measure of teachers' knowledge was used, rather than proxy measures such as mathematical qualifications.

There were a number of differences between the *StatSmart* study and those of Hill et al. (2005) and Baumert et al. (2010). These two studies identified a link between teachers' PCK and their current students' learning outcomes. *StatSmart*, in contrast, was a 3-year longitudinal study in which the context of the project meant that there was no control by researchers over which grades and classes teachers taught, at what point in the school year statistics was taught or any changes to teachers and classes during the study. During the project, students changed classes and teachers, sometimes into classes taught by teachers not participating in *StatSmart*, some teachers left their schools and others taught different grades from year to year. In addition, as is common in the Australian situation, where students were grouped by ability, teachers taught different ability groups of students from year to year and, sometimes, in the same year had a high- and low-ability group, both of which undertook the *StatSmart* tests. In this naturalistic situation, untangling the influence of a particular current teacher proved impossible because of the number of uncontrolled variables.

PCK could be considered as a measure of teacher quality, with teachers having higher levels of PCK being more likely to teach classes showing gains in content knowledge (Baumert et al., 2010; Hill et al., 2004, 2005). The study design was, therefore, a pre- and post-test followed by a follow-up test to see to what extent changes in students' outcomes were maintained (Callingham & Watson, 2007). This longitudinal design allowed a consideration of both the influence of the initial teachers' PCK on students' outcomes when they were actively teaching the class, and also the maintenance of that influence over time.

A decision was thus made to consider only the initial teacher's PCK on a particular student's outcomes. It is known that students' prior achievement is a predictor of future learning outcomes (Dochy et al., 1999), and this was reflected in the study design.

Hence, it is not unreasonable to suppose that the influence of an initial teacher's PCK might continue to impact on students' future learning outcomes.

Methodology

The Context

The context of the research reported here was a 3-year research project, *StatSmart*, in conjunction with the Australian Bureau of Statistics (ABS), the US manufacturer of the software *Fathom* (Finzer, 2002) and *TinkerPlots* (Konold & Miller, 2005) and an independent expert in professional learning for teachers of mathematics. Initially, 42 teachers were chosen from 18 schools in three Australian states. A commitment was made by the teachers and schools to implement statistics units within the middle school years (grades 5 to 9) and their state's mathematics curriculum, based on the research findings on the development of student understanding (Watson, 2006) and employing one or both of the software packages that were provided to every school. To assist further, each year, a 2-day workshop with all expenses paid was held in the ABS offices in Melbourne, including the software developers from the USA. Teachers were expected to complete a teacher profile, including items developed to measure PCK. Examples of items used in the profile are found in Watson, Callingham & Donne (2008) based on proportional reasoning and in Callingham & Watson (2011) based on odds.

Participants

Students

Students in the middle years of schooling (ages 10 to 15 years) together with their teachers were the target groups. These students and teachers were located in three Australian states that had similar but not identical curricula (see Callingham, 2010 for details). Over the course of the project, each student undertook three surveys of statistical literacy. The first two were taken at the start and end of the first year in which they entered the study, and the third survey was a follow-up taken about 12 months after the second survey. Over the 3-year study, there were two phases of students who completed all three surveys and one phase that completed only the first two surveys (see Callingham & Watson, 2007 for details of the research design). A small number of students completed a fourth survey because they happened to be in a class taught by a project teacher at the time the survey was completed.

The sample used in the analysis reported here consisted of 789 students for whom there were three or four data points over 3 years and who did not change schools. All of these students were part of phase 1 and phase 2 of the study. More specifically, 70 students had four observations and the remaining 719 had three observations. All of the students did an initial survey and a follow-up survey after about 6 months while they were still in the same year groups at school. After 12 months, most of them ($n=765$) completed a longitudinal survey. A small number had different participation patterns (see Table 1 for details). All of these students were taught by teachers who had completed the initial teacher survey and for whom there were PCK measures available.

Table 1 Test participation patterns for 789 students with teacher PCK data

| Rd 1 | Rd 2 | Rd 3 | Rd 4 | Rd 5 | Rd 6 | Total |
|------|------|------|------|------|-------|-------|
| √ | √ | | √ | | √ | 70 |
| √ | √ | | √ | | | 352 |
| √ | √ | | | | √ | 24 |
| | | √ | √ | | √ | 343 |
| | | | | | Total | 789 |

Rd 1, Rd 2, etc. refer to the *StatSmart* test rounds

There was an approximately even split of male and female students (48.7 % male), and just over 10 % of the students came from backgrounds where they did not speak English at home (non-English speaking background (NESB)=10.6 %). When they commenced the study, students' ages ranged from 10.1 to 15.8 years ($M=12.9$, $SD=1.0$), and most (72 %) of them were attending a secondary school.

Teachers

At their first test, these students were taught by 36 different teachers located in 15 schools. Of these teachers, just over half (56 %) were male. Teachers also completed three surveys, one in each year of the project. Particular care was taken to ensure that teachers could be associated with particular groups of students in order to associate teachers' measured knowledge directly with students' outcomes. The teachers had varied backgrounds in both the level of mathematics studied and teaching experience, summarised in Table 2.

Instruments and Analysis

Surveys

Teachers completed a profile instrument that included a set of 12 items designed to measure PCK in statistics, rather than general mathematics. The PCK items were based on real students' survey responses from previous studies to provide authenticity.

Table 2 Characteristics of *StatSmart* teachers

| Mathematics background | Number of teachers | Mathematics teaching experience | Number of teachers | |
|------------------------|--------------------|---------------------------------|--------------------|------|
| No maths | 3 | 91 % <2 years | 0 | |
| 1 semester tertiary | 8 | 23 % 2–5 years | 5 | 14 % |
| 1 year tertiary | 10 | 28 % 6–10 years | 0 | |
| Undergraduate major | 14 | 40 % 11–15 years | 10 | 29 % |
| | | 16–25 years | 7 | 20 % |
| | | >25 years | 13 | 37 % |

Figure 1 contains an example of an item that had been used in student surveys and was intended to measure teachers' content knowledge, knowledge of students as learners and pedagogical content knowledge for intervention in the classroom. Contextually, it refers to shops common in Australia and had been used in several previous studies. For this item, students' actual responses provided a pool of examples of students' thinking (e.g. Watson & Callingham, 2003) against which teachers' responses could be compared and coded.

A second item type asked teachers to respond to students' actual answers to survey questions (see Fig. 2 for an example). These items addressed teachers' capacity to provide student-centred ideas for intervention. Scoring rubrics were developed for teachers' responses to both types of survey question based on increasing complexity and mathematisation. The rubrics for questions 5.3 and 5.4 are also shown in Fig. 2.

Students undertook one of three test forms, all linked by a core of 10 common items (Callingham & Watson, 2007). Each test had between 22 and 24 items addressing different aspects of statistical literacy, hereafter termed statistical literacy knowledge (SLK), including measures of central tendency, variation and sampling. The student tests included several items that were also answered by teachers, including the supermarket item shown in Fig. 1, and items addressing two-way tables (Watson & Callingham, 2014), and likelihood and sample size (Watson & Callingham, 2013). Hence, the student and teacher surveys addressed the same content, albeit from different perspectives.

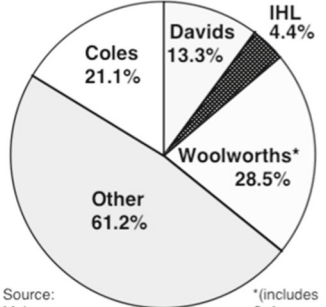
Both student and teacher responses to the surveys were analysed using Rasch measurement using the software Winsteps 3.75.0 (Linacre, 2012). Rasch analysis (Bond & Fox, 2007; Rasch, 1960) uses the interactions between items and test takers (persons) to place all items and persons on the same interval scale. The approach is

Coles Myer accelerates retail purge

What is the newspaper trying to tell its readers in this pie chart?

Is there anything unusual about it?

Nationwide retail grocery market shares



| Retailer | Market Share |
|-------------|--------------|
| Other | 61.2% |
| Woolworths* | 28.5% |
| Coles | 21.1% |
| Davids | 13.3% |
| IHL | 4.4% |

Source: McLennan Magasnik Associates

*(includes Safeways in Victoria)

4.2a
What responses would you expect from your students? Write down some appropriate and inappropriate responses (use * to show appropriate responses).

4.2b
How would/could you use this item in the classroom? For example, how would you intervene to address the inappropriate responses?

Fig. 1 Typical teacher PCK item addressing knowledge of students and knowledge of pedagogy

Consider the following problem that students were asked in a survey about chance and data:

The average number of children in 10 families in the neighbourhood is 2.3.
 One family with 5 children leaves the neighbourhood. What is the average number of children per family now?
 Show your work here.

Consider each of the following answers and explanations given by students in response to the problem.
Explain how you would respond to each answer.

5.3
 $2.3 \times 10 = 23 - 5 = 18 \div 10 = 1.8$

| Code | Description |
|------|---|
| 1 | General response not involving the mathematics of the problem: "get student to explain thinking." |
| 2 | Comment on number of families or equation structure (problem content only). |
| 3 | Questioning of student in relation to one of the issues: number of families <u>or</u> equation structure. |
| 4 | Sequencing of task with questions for student to complete. |

5.4
 I don't know how many children in each family so how do you work it out?

| Code | Description |
|------|--|
| 0 | Unsure how to proceed/no mathematics in response. |
| 1 | Single isolated question or suggested approach, e.g., discuss average and how to work out. |
| 2 | Extended explanation related to formulas involved. |
| 3 | Suggestions that go beyond the formula to model the problem. |

Fig. 2 An example of student-answered items, together with the scoring rubrics, used in the teacher survey

based on a probabilistic model underpinned by three key assumptions: (i) the items address a single unidimensional construct; (ii) the probability of a correct or higher level response increases monotonically with an increase in a person's ability or understanding; and (iii) all items are independent of each other. Where these assumptions are violated, the fit to the model falls outside acceptable parameters. Hence, the fit to the model becomes of prime importance in determining the validity of the construct and the suitability of the measures obtained for the intended purpose. The specific model used was the Partial Credit Model (PCM) (Masters, 1982) where the scoring rubrics, for both students and teachers, were used to provide partial scores.

Model fit is reported by Rasch modelling programs as four statistics: the Infit is a weighted least squares measure and the Outfit is the unweighted measure. Both have an ideal value of 1.0, and values suitable for measurement purposes lie between 0.5 and 1.5 (Linacre, 2002). In addition, a standardised *z* score is provided for each with acceptable values lying between ±2. Rasch reliability statistics are the item and person separation indices. These provide a measure of the consistency with which persons or items are located on the scale produced. In general, person separation is considered satisfactory if the index is >0.8 and item separation is satisfactory if the index is >0.9 (Linacre, 2013). Both indices are uninfluenced by model fit. The fit to the model and reliability indices for the tests used in the analysis reported here for both students' SLK and teachers' PCK are summarised in Table 3. Item separation indices are not available for tests that are anchored to a previous administration. All fit and reliability statistics for the tests used in the analysis were generally acceptable.

Rasch person measures in logits, the logarithm of the odds of success used as the unit of Rasch measurement, were estimated for each of the three student tests. These estimates were anchored to the first test to ensure that all were directly comparable on the same measurement scale (Bond & Fox, 2007). A range of demographic variables

Table 3 Summary statistics for item (I) and person (P) measures for student SLK tests and teacher PCK assessment

| Test | Rasch item separation index | Rasch person separation index | Infit (I) | Infitz (I) | Outfit (I) | Outfitz (I) | Infit (P) | Infitz (P) | Outfit (P) | Outfitz (P) |
|-------------|-----------------------------|-------------------------------|-----------|------------|------------|-------------|-----------|------------|------------|-------------|
| SLK 1 | 0.99 | 0.86 | 0.99 | -0.20 | 0.98 | -0.30 | 1.06 | 0.10 | 0.98 | 0.00 |
| SLK 2 | Anchored | 0.85 | 1.06 | 0.70 | 1.08 | 0.70 | 1.13 | 0.40 | 1.13 | 0.50 |
| SLK 3 | Anchored | 0.84 | 1.10 | 1.20 | 1.11 | 1.20 | 1.17 | 0.60 | 1.18 | 0.70 |
| Teacher PCK | 0.93 | 0.77 | 1.00 | 0.10 | 0.99 | 0.00 | 1.03 | 0.00 | 0.99 | 0.00 |

was also included. For the purpose of the analysis reported here, only the PCK measure applicable to the students' first test was used, usually the teacher's initial measure. These measures were then used as input variables to create hierarchical models.

Analysis

Simple descriptive techniques were initially used to explore the impact of teacher factors on measures of students' statistical understanding. Multilevel regression models were then used to control for the effects of demographic variables. These models were used to capture the longitudinal nature of the outcome variable and dependence between students attending given schools. It was not possible to model the dependence of students within classes, because nearly two thirds of the students (63 %) changed teachers after the second test, often to teachers who were not part of the project, and from whom there were no PCK measures available. In addition, only students who had undertaken test 1 or test 2 were tracked, so that classes taught by a non-*StatSmart* teacher did not provide intact class data. Model estimates were obtained using the software package R (R Development Core Team, 2011) and in particular the Multilevel package (Bliese, 2012) as described in Faraway (2006).

Results

Descriptive Analysis

Initial results considered the changes in the overall SLK scores across time. At the student level, SLK scores in tests 1 and 2 were correlated ($r(787)=.67, p<.01$), as were scores between tests 2 and 3 ($r(787)=.67, p<.01$). On average, students obtained relatively low SLK scores in their first test ($M=-0.53, SD=0.71$) and these improved in their second test ($M=-0.25, SD=0.64$), with this increase statistically significant ($t(1521)^1=9.4, p<.01, d\approx 0.4$). The scores, however, appeared to decline slightly for the third test ($M=-0.27, SD=0.65$), though this was not statistically significant ($t(1547)=1.4, p<.01, d\approx 0.05$)². This pattern is not unexpected in studies of this type, where the pre- and post-tests occur within a relatively short period, and the longitudinal test is some considerable time after the post-test

¹ Degrees of freedom based on Welch's *t* test

² The 70 students who did four tests reported a non-significant increase of 0.05 logits on their last test.

(Cohen, Manion & Morrison, 2011). The growth pattern also appeared to be influenced by grade level with differences occurring between test 2 and test 3. Figure 3 shows the comparative results in three tests for two pairs of grade groupings, Grades 5 and 6, and Grades 8 and 9 as box-and-whisker plots. The box shows the inter-quartile range of scores, and the heavy line across each box is the median. The small circles represent scores falling below the 10th percentile or above the 90th percentile. As shown in Fig. 3, the 124 students in grades 5 and 6 showed a slight increase between these tests, whereas the scores of the 285 students in grades 8 and 9 fell slightly. Similar patterns have been shown in other studies across the middle years of schooling (e.g. Hill, Rowe, Holmes-Smith & Russell, 1996).

Next, the impact of the teacher on students' SLK scores was explored. Given the restriction caused by changing teachers within schools, comparisons were made at the student rather than teacher level. PCK scores for the initial teacher, the one teaching the students for the period including tests 1 and 2, ranged from -1.61 to 2.47 logits ($M=0.25$, $SD=0.78$). These were weakly correlated with the students' SLK scores in tests 1 ($r(787)=-.17$, $p<.01$) and 2 ($r(787)=.08$, $p=.02$), but not in test 3 ($r(787)=.00$, $p=.80$), suggesting a waning effect. By way of comparison, the correlation between teachers' PCK scores and students' SLK scores for test 1, aggregated to the teacher level, revealed a similar association ($r(34)=.27$, $p=.11$), suggesting the reported positive association is not an artefact of the grouping in these data.

Given the large number of students who changed teachers between tests 2 and 3, the effect due to this change was also considered. In particular, the results of students who did not change teachers were compared with those who changed to a teacher in the *StatSmart* project and with those who changed to a teacher not in the *StatSmart* project (see Table 4). As is seen in the table, the students who changed to a non-*StatSmart* teacher tended to have a fall in results between tests 2 and 3, though this was only significant at the 10 % level ($t(977)=1.7$, $p=.08$, $d\approx 0.1$). Students who did not change teachers and those who changed to *StatSmart* teachers experienced small, non-significant gains. Those who did not change teachers started from a much lower mean

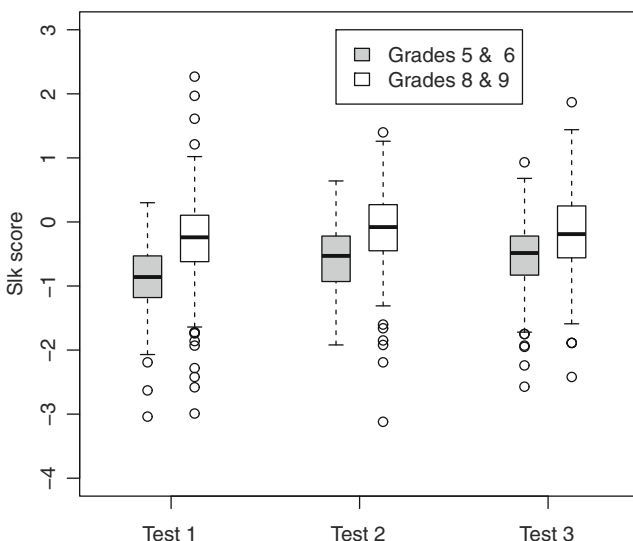


Fig. 3 Distribution of SLK scores across three tests for younger and older students

Table 4 Influence of changing teachers on students' SLK scores

| Status | Mean/SD test 1 | Mean/SD test 2 | Mean/SD test 3 | Number of students |
|--|----------------|----------------|----------------|--------------------|
| Changed to non- <i>StatSmart</i> teacher | -0.44/0.64 | -0.18/0.56 | -0.24/0.64 | 498 |
| Changed to <i>StatSmart</i> teacher | -0.91/0.89 | -0.54/0.75 | -0.47/0.67 | 125 |
| Did not change teachers | -0.52/0.66 | -0.23/0.70 | -0.20/0.62 | 166 |

result in their first test than the other two groups. This particular group tended to be younger ($M=12.6$ years) than the other students ($M=12.9$ years), and as is seen in Fig. 1, younger students tended to score lower in these tests than their older peers. They were also more likely to be in a primary school setting, where teachers may follow a class of students over more than 1 year. Continuing to be taught by a teacher who was involved with the project, however, appeared to have a positive influence on students' outcomes, and this finding is further considered in the [Discussion](#).

Other teacher factors were also considered. Students taught by female teachers, for example, on average scored lower than students taught by male teachers in test 1 ($t(771)=6.3, p<.01, d\approx 0.4$), test 2 ($t(782)=5.8, p<.01, d\approx 0.4$) and test 3 ($t(787)=5.0, p<.01, d\approx 0.3$). Further, those students taught by teachers with a tertiary-level mathematics background performed better than those taught by less-qualified teachers in each of test 1 ($F(3, 776)=20.6, p<.01$), test 2 ($F(3, 776)=18.0, p<.01$) and test 3 ($F(3, 776)=6.6, p<.01$). In both cases, however, the lower performing groups were likely to be younger students in primary settings.

One school-level factor was also considered. The Indicator of Community Socio-Educational Advantage (ICSEA) measure (Australian Curriculum, Assessment and Reporting Authority, 2012) was used to assess the impact of school-level socioeconomic status on students' SLK scores. The mean for ICSEA is set at 1000, and the schools in the study had ICSEA scores ranging from 912 to 1168. At the school level, correlations between this index and mean SLK scores were positively associated in each of test 1 ($r(13)=.65, p<.01$), test 2 ($r(13)=.67, p<.01$) and test 3 ($r(13)=.69, p<.01$), suggesting that ICSEA has a strong influence on between-school variation.

Multivariate Analysis

Initially, an analysis of variance was conducted on the SLK results for each of the three tests against class and school groups. For test 1, results suggested that 15 % of the variance could be attributed to between-school effects, 28 % to between-class effects and the remainder to between-student effects. Similar results were obtained for the other tests, suggesting that grouping by classes and schools was desirable. As reported earlier, however, grouping by classes was not possible because most students changed teachers after test 2, a situation common in the Australian context. Instead, the change of teachers was modelled using a change-teacher variable. Given the risk that standard errors may be overestimated, a more stringent critical value of 1 % was adopted, as recommended by Thomas (2001).

A random intercept model was applied to these data. The model assumes that at the individual student level, growth in statistical understanding is linear and expressed as

$$Y_{ij} = \beta_{0ij} + \beta_1 t + \epsilon_{ij} \quad (1)$$

where the errors (ϵ_{ij}) are assumed to be independent, distributed normally and with a common variance. Y_{ij} is the mathematics achievement of student i , from school j , at time t ($t=0, 0.4, 1.4$, and for some 2.4 years). The initial status of student i from school j is denoted β_{0ij} , and the model assumes that growth during the period of the study is the same³ for each student β_1 . The parameter β_{0ij} , however, is assumed to vary randomly across students within schools, in that

$$\beta_{0ij} = \gamma_{00j} + u_{0ij} \quad (2)$$

where γ_{00j} is the estimated initial mean score for all students attending school j and u_{0ij} is the discrepancy between this and the initial score of student i in school j . The parameter γ_{00j} is assumed to vary randomly across schools, in that

$$\gamma_{00j} = \tau_{000} + u_j \quad (3)$$

where τ_{000} is the grand mean initial score for all students across all schools and u_j the discrepancy between this grand mean and the mean for school j .

Equations (1), (2) and (3) above represent the unconditional model reported as model 1 in Table 5. In line with recommendations from Dedrick et al. (2009), underlying assumptions of the model, including the absence of an autoregressive structure, were assessed and found to be satisfactory. As is seen in Table 3, the null model predicted that students' participation in the study for 1 year was associated with an increase in SLK of 0.14 logits.

In developing the final model, several factors were introduced in order to explain each of the variance components in the null model: the residual or within-student variance, $\text{var}(\epsilon_{ij})$; the between-student variance, $\text{var}(u_{0ij})$; and the between-school variance, $\text{var}(u_j)$. Changing teachers, for example, was found to be a significant predictor of SLK scores that reduced within-student variance from 0.134 to 0.128 (5 % reduction). Student-level factors, including their standardised age when they completed test 1 (agez), whether they spoke a language other than English (NESB=1) and the standardised PCK score of their initial teacher (PCKZ), were significant predictors of SLK that reduced between-student variance from 0.185 to 0.178 (6 % reduction). Other teacher factors such as sex and mathematical background did not predict SLK in the model. Finally, the standardised ICSEA index (ICSEAZ) was found to be a significant predictor of SLK that reduced between-school variance from 0.089 to 0.028 (69 % reduction).

Given the results of the descriptive analysis, two interactions were then introduced into the model. The first was an interaction between PCK and time, in that the earlier analysis suggested a waning effect. The second was an interaction between commencement age in the project and time, in that results displayed in Fig. 1 suggest that older students typically had greater initial scores than their younger peers but less steep growth trajectories. Both of these interactions were found to be significant predictors of SLK that improved model fit (based on a comparison of deviance test). The final model is shown as model 2 in Table 5.

As is seen in model 2 of Table 5, changing teacher to a *StatSmart* teacher (chgtss=1) was not significantly different to not changing teachers at all. Changing teacher to a

³ A random slopes model that allows different growth trajectories was also tested but failed to converge.

Table 5 Results of multilevel models

| Variable | Model 1 | | Model 2 | |
|---------------------------------------|--------------------|------|----------|-------------------|
| | Estimate | SE | Estimate | SE |
| Fixed effects | | | | |
| Initial mean SLK (γ_{00}) | -0.44 ^a | 0.08 | -0.42 | 0.05 |
| Time | 0.14 | 0.01 | 0.20 | 0.02 |
| chgtss | | | -0.06 | 0.04 ^b |
| chgtns | | | -0.14 | 0.03 |
| agez | | | 0.16 | 0.02 |
| NESB | | | -0.17 | 0.06 |
| PCKZ | | | 0.07 | 0.02 |
| ICSEAZ | | | 0.23 | 0.04 |
| PCKZ * time | | | -0.03 | 0.01 |
| agez * time | | | -0.05 | 0.01 |
| Random effects | | | | |
| Within-student var(ϵ_{it}) | 0.134 | | 0.128 | |
| Between-student var(u_{0i}) | 0.185 | | 0.173 | |
| Between-school var(u_{1i}) | 0.089 | | 0.028 | |
| Model deviance | 3354 | | 3205 | |
| Number of parameters | 5 | | 13 | |

^a All effects are in logits

^b This effect is not statistically significant. All others are significant at the 1 % level

non-*StatSmart* teacher (chgtns=1), however, was associated with a significant reduction in SLK scores ($\gamma = -0.14$). Student-level factors, including their standardised age when they completed test 1 (agez), whether they spoke a language other than English (NESB=1) and the standardised PCK score of their initial teacher (PCKZ), were significant predictors of SLK that reduced between-student variance. Students with non-English-speaking backgrounds, for example, were predicted to score 0.17 logits lower than their peers throughout the study. The model also suggests that teachers' initial PCK was associated with higher SLK scores ($\gamma = 0.07$), but that this association fell by 0.03 logits with each year that the student was in the study. The ICSEA index was found to be a significant predictor of SLK that reduced between-school variance, in that students from schools with an ICSEA value one standard deviation higher than the mean were predicted to score on average 0.23 logits higher than their peers.

Discussion

In the final model, a substantial part of the between-school variance was explained by socio-economic factors represented by the school variable ICSEA. Age of the student at the first test and whether the student spoke a language other than English at home

contributed substantially to between-student variance explained. These results are not surprising and echo those from other studies (Hattie, 2008). Of particular interest, however, is the effect of the first teacher's measured pedagogical content knowledge (PCK). This variable had a significant effect on students' achievement, in line with other studies undertaken with elementary teachers (Hill et al., 2005), or with students nearing the end of high school (Baumert et al., 2010). The *StatSmart* study addressed the middle years of schooling, hence establishing that PCK is a key variable for considering teachers' impacts on their students' learning across the years of schooling. The PCK/time interaction term, however, was negative, suggesting a waning effect, in that the influence of good, or bad, teachers wanes as students progress through school. Intuitively, this finding seems sensible. It is the current teacher who is likely to have immediate impact, but because of the nature of the study, modelling this effect proved impossible. The teacher's mathematical background, gender and years of teaching experience had no significant impact on students' measured achievement, in line with other studies (e.g. Mewborn, 2001). Although this study was undertaken in the context of statistics education, there is no reason to suppose that it would not apply in the mathematics domain more generally, given that statistics is taught within the mathematics curriculum.

StatSmart was able to track both students and their teachers across time. It proved difficult to find similar studies in which both student and teacher achievement data were available and linked together, other than the two referred to earlier (Hill et al., 2005; Baumert et al., 2010). Sustainability across time appears to be imperative at the teacher level. Effective schools are known to provide consistency for students as they move up the grade levels (Hill et al., 1996; Hill & Rowe, 1998). The finding that changing to a non-*StatSmart* teacher had a negative effect on students' measured achievement is important. Much is made in the research literature of key transitions, such as the move from primary to secondary school, but there is little to identify other transitions.

The *StatSmart* study appears to indicate that moving from one teacher to another teacher having similar professional learning experiences reduced any negative effects of transition. Rowan et al. (2002) argued that using students' individual growth trajectories tracked across at least three time points, as was done in this study, is preferable to using single achievement scores or gain scores. They also showed that when elementary students moved from class to class across years, the effect of changing teachers was inconsistent, with some students making gains and others not. The difficulties of establishing teacher effects across more than one class are well documented (e.g. Hill & Rowe, 1998), especially in systems where schools attempt to create classes each year that take account of the individual student's needs, as is common in Australia. The finding reported here has potential implications for systems and schools. Although the identified teacher effects were small, they are educationally important. Teacher quality is likely to be more amenable to policy intervention than are the large effect variables of socio-economic status (ICSEA) and non-English-speaking background (NESB) (Hattie, 2008).

Time in the programme also had a relatively large effect on students' SLK scores. Partly this is explained by increasing age and experience. The finding, however, has potential implications for both policy makers and schools when taken together with the changing teacher effects. Having a sustained focus by teachers who participated in

professional learning for an extended period had a positive effect on their students' achievement in this study. Many professional learning programmes are undertaken by schools or systems for short periods, in line with funding availability. The *StatSmart* project was a 3-year programme and retained a majority of the original teachers for the whole period. These teachers were, by their continued presence in the project, highly committed and during the 3 years made changes to their practice, and reported back on these at annual conferences. In addition, there was ongoing contact with the research team. Only rarely do projects such as *StatSmart* hold teachers over time, especially where the schools are "conscripted" into professional learning studies by the funding bodies (e.g. Watson, Brown, Beswick & Wright, 2011). The implication is that education systems and schools need to make a long-term commitment to a particular programme or approach to teaching, rather than commonly occurring situations where one-off professional learning is delivered by an expert through workshops disconnected from teachers' classrooms.

The findings from this study must be considered in the light of the limitations imposed by the naturalistic setting. There are myriad uncontrolled variables that impact on students, in classes, in schools. These create considerable "noise" and unexplained variance in the data collected, and this is acknowledged. Nevertheless, the findings are similar to those of other studies conducted in more controlled conditions. In addition, the loss of teachers and students from the study over time meant that the data set comprising a complete set of both teachers' and students' measures became too small to achieve sufficient statistical power. If funding had permitted a much larger data set at the start of the study, the study design meant that it might have been possible to track changes in teachers' and students' outcomes across time and the association between these changes. Within the constraints of the *StatSmart* study, this proved impossible.

Conclusion

This study has indicated that teachers' pedagogical content knowledge in statistics was associated with their students' learning outcomes in different educational contexts to those reported in previous studies. In addition, the findings indicated that negative effects due to transitions to new teachers can be mitigated if the new teacher has similar professional learning experiences. With these two findings, the *StatSmart* study has added to the growing body of evidence that knowing the subject matter alone is not sufficient for positive teaching outcomes. It is the specialised way in which teachers understand their subject that counts. The complex blend of subject matter knowledge and understanding of student learning and school context, known as pedagogical content knowledge, makes a difference, together with a sustained focus on, in this instance, specific professional learning.

One next step is to consider whether particular groups of students, such as low achievers, benefit more than others. Another is to consider how the nature of PCK changes with levels of schooling, and the effects of different approaches to developing teachers' PCK. There is still much to be researched in the area of teacher knowledge and its influence on student outcomes.

Acknowledgments This project was funded by Australian Research Grant No. LP0669106 in collaboration with the Australian Bureau of Statistics, Key Curriculum Press and The Baker Centre for School Mathematics, Prince Alfred College, Adelaide, South Australia.

References

- Australian Education Council (1991). *A national statement on mathematics for Australian schools*. Melbourne, Australia: Author.
- Australian Curriculum, Assessment and Reporting Authority. (2011). *The Australian curriculum*. Sydney, Australia: Author. Retrieved from <http://www.australiancurriculum.edu.au/>. Accessed 27 May 2015.
- Australian Curriculum, Assessment and Reporting Authority. (2012). *Guide to understanding ICSEA*. Sydney, Australia: Author. Retrieved from http://www.acara.edu.au/verve/_resources/Guide_to_understanding_ICSEA.pdf. Accessed 27 May 2015.
- Bliese, P. (2012). *Multilevel modelling in R (2.5)*. Columbia, SC: University of South Carolina. Available online at http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf. Accessed 27 May 2015.
- Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.
- Callingham, R. (2010). Trajectories of learning in middle years' students' statistical development. In C. Reading (Ed.) *Data and context in statistics education: Towards an evidence-based society* (Proceedings of the 8th International Conference on the Teaching of Statistics, Ljubljana, Slovenia, July)[CDRom]. Voorburg, The Netherlands: International Statistical Institute.
- Callingham, R. & Watson, J.M. (2007). Overcoming research design issues using Rasch measurement: The StatSmart project. In P. Jeffery (Ed.) *Proceedings of the AARE annual conference, Fremantle, 2007*. Available at <http://www.aare.edu.au/07pap/cal07042.pdf>.
- Callingham, R. & Watson, J.M. (2011). Measuring levels of statistical pedagogical content knowledge. In C. Batanero, G. Burrill & C. Reading (Eds.) *Teaching statistics in school mathematics—challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 283–293). Dordrecht, The Netherlands: Springer.
- Callingham, R., Watson, J. & Burgess, T. (2012). Uncertainty in mathematics education: What to do with statistics? In J. Greenlees, T. Logan, T. Lowrie, A. MacDonald & B. Perry (Eds.), *Review of Australasian mathematics education research: 2008–2011*. Rotterdam, The Netherlands: Sense.
- Ball, D. L., Thames, M. H. & Phelps, G. (2008). Content knowledge for teaching: What makes it so special? *Journal of Teacher Education*, 59(5), 389–407.
- Batanero, C., Estepa, A., Godino, J. D. & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education*, 27, 151–169.
- Batanero, C., Burrill, G. & Reading, C. (2011). *Teaching statistics in school mathematics—challenges for teaching and teacher education: A joint ICMI/IASE study*. Dordrecht, The Netherlands: Springer.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... Yi-Miau, T. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180.
- Beswick, K., Callingham, R. & Watson, J.M. (2012). The nature and development of middle school mathematics teachers' knowledge. *Journal of Mathematics Teacher Education*, 15(2), 131–157. doi:10.1007/s10857-011-9177-9.
- Cai, J. (1995). Beyond the computational algorithm: Students' understanding of the arithmetic average concept. In L. Meira & D. Carraher (Eds.), *Proceedings of the 19th psychology of mathematics education conference* (Vol. 3, pp. 144–151). São Paulo, Brazil: PME Program Committee.
- Cai, J. (1998). Exploring students' conceptual understanding of the averaging algorithm. *School Science and Mathematics*, 98, 93–98.
- Chick, H.L. (2007). Teaching and learning by example. In J. M. Watson & K. Beswick (Eds.), *Mathematics: Essential research, essential practice. Proceedings of the 30th Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 3–21). Sydney, Australia: MERGA.
- Cohen, L., Manion, L. & Morrison, K. (2011). *Research methods in education*. Abingdon, England: Routledge.
- Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Dromrey, J. D., Lang, T. R., ... Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, 79(1), 69–102.

- Dochy, F., Segers, M. & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research*, 69(2), 145–186.
- Even, R. & Tirosh, D. (2002). Teacher knowledge and understanding of students' mathematical learning. In L. English (Ed.), *Handbook of international research in mathematics education* (pp. 219–240). Mahwah, NJ: Erlbaum.
- Faraway, J. J. (2006). Extending the linear model with R. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4), 1008.
- Finzer, W. (2002). *Fathom dynamic data software (version 2.1)*. [computer software]. Emeryville, CA: Key Curriculum Press.
- Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M. & Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) report: a Pre-K–12 curriculum framework*. Alexandria, VA: American Statistical Association. Retrieved from <http://www.amstat.org/education/gaise/>
- Green, D. R. (1983). A survey of probability concepts in 3000 pupils aged 11–16 years. In D. R. Grey, P. Holmes, V. Barnett & G. M. Constable (Eds.), *Proceedings of the first international conference on teaching statistics* (Vol. 2, pp. 766–783). Sheffield, England: Teaching Statistics Trust.
- Green, D.R. (1986). Children's understanding of randomness: Report of a survey of 1600 children aged 7–11 years. In R. Davidson & J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics* (pp. 287–291). Victoria, BC: The Organizing Committee, ICOTS2.
- Green, D. (1991). A longitudinal study of pupils' probability concepts. In D. Vere-Jones (Ed.) *Proceedings of the Third International Conference on Teaching Statistics. Vol. 1. School and general issues* (pp. 320–328). Voorburg, The Netherlands: International Statistical Institute.
- Groth, R. E. (2007). Toward a conceptualisation of statistical knowledge for teaching. *Journal for Research in Mathematics Education*, 38, 427–437.
- Hattie, J. (2008). *Visible learning*. London, England: Routledge.
- Hill, P. W. & Rowe, K. J. (1998). Modelling student progress in studies of educational effectiveness. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, 9(3), 310–333. doi:10.1080/0924345980090303.
- Hill, P. W., Rowe, K. J., Holmes-Smith, P. & Russell, V. J. (1996). *The Victorian Quality Schools Project: A study of school and teacher effectiveness. Report to the Australian Research Council (Volume 1 - Report)*. Melbourne, Australia: Centre for Applied Educational Research, Faculty of Education, The University of Melbourne.
- Hill, H. C., Schilling, S. G. & Ball, D. L. (2004). Developing measures of teachers' mathematics for teaching. *Elementary School Journal*, 105, 11–30.
- 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.
- Hill, H. C., Sleep, L., Lewis, J. M. & Ball, D. L. (2007). Assessing teachers mathematical knowledge: What knowledge matters and what evidence counts? In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 111–156). Reston, VA: National Council of Teachers of Mathematics.
- Konold, C. & Higgins, T. L. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics.
- Konold, C. & Miller, C. D. (2005). *Tinkerplots: Dynamic data exploration*. [Computer software] Emeryville, CA: Key Curriculum Press.
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 278.
- Linacre, J. M. (2012). *Winsteps Rasch measurement 3.75.0* [computer software]. Retrieved from Winsteps.com. Accessed 27 May 2015.
- Linacre, J. M. (2013). *A users' guide to Winsteps*. Retrieved from <http://Winsteps.com>. Accessed 27 May 2015.
- Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Mahwah, NJ: Lawrence Erlbaum.
- Makar, K. & Fielding-Wells, J. (2011). Teaching teachers to teach statistical investigations. In C. Batanero, G. Burrill & C. Reading (Eds.), *Teaching statistics in school mathematics—challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 347–358). Dordrecht, The Netherlands: Springer.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 49, 359–381.
- Mevarech, Z. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics*, 14, 415–429.

- Mewborn, D. S. (2001). Teachers' content knowledge, teacher education, and their effects on the preparation of elementary teachers in the United States. *Mathematics Teacher Education and Development*, 3, 28–36.
- Ministry of Education. (1992). *Mathematics in the New Zealand curriculum*. Wellington, New Zealand: Author.
- Ministry of Education. (2007). *Draft Mathematics and Statistics Curriculum*. Wellington, New Zealand: Author.
- Mokros, J. & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, 26, 20–39.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
- Pierce, R. & Chick, H. (2011). Teachers' beliefs about statistics education. In C. Batanero, G. Burrill & C. Reading (Eds.), *Teaching statistics in school mathematics—challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 151–162). Dordrecht, The Netherlands: Springer.
- Pollatsek, A., Lima, S. & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics*, 12, 191–204.
- R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna, Austria: The R Foundation for Statistical Computing. Available at <http://www.R-project.org/>. Accessed 27 May 2015.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.
- Rowan, B., Correnti, R. & Miller, R.J. (2002). *What large-scale, survey research tells us about teacher effects on student achievement: Insights from the 'Prospects' study of elementary schools*. (CPRE Research Report Series RR-051). Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania. Available at http://www.cpre.org/images/stories/cpre_pdfs/rr51.pdf.
- Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester Jr. (Ed.), *Second handbook on research on mathematics teaching and learning* (pp. 957–1009). Charlotte, CA: Information Age Publishing.
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57, 1–22.
- Strauss, S. & Bichler, E. (1988). The development of children's concept of the arithmetic average. *Journal for Research in Mathematics Education*, 19, 64–80.
- Thomas, S. L. & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education*, 42(5), 517–540.
- 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. M. (2006). *Statistical literacy at school: Growth and goals*. Mahwah, NJ: Erlbaum.
- Watson, J. M. & Moritz, J.B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, 37, 145–168.
- Watson, J. M. & Moritz, J.B. (2000). Developing concepts of sampling. *Journal for Research in Mathematics Education*, 31, 44–70.
- Watson, J. M. & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3–46.
- Watson, J. M. & Callingham, R. (2013). Likelihood and sample size: The understandings of students and their teachers. *Journal of Mathematical Behaviour*, 32(3), 600–672.
- Watson, J. M. & Callingham, R. (2014). Two-way tables: Issues at the heart of statistics and probability for students and teachers. *Mathematical Thinking and Learning*, 16(4), 254–284.
- Watson, J. M., Callingham, R. & Kelly, B.A. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning*, 9, 83–130.
- Watson, J. M., Callingham, R. & Donne, J. (2008). Proportional reasoning: Student knowledge and teachers' pedagogical content knowledge. In M. Goos, R. Brown & K. Makar (Eds.), *Navigating currents and charting directions* (Proceedings of the 31st annual conference of the Mathematics Education Research Group of Australasia, Brisbane, Vol. 2, pp. 555–562). Adelaide, Australia: MERGA.
- Watson, J. M., Brown, N., Beswick, K. & Wright, S. (2011). Teacher change in a changing educational environment. In J. Clark, B. Kissanee, J. Mousley, T. Spencer & S. Thornton (Eds.), *Mathematics: Traditions and [new] practices* (Proceedings of the AAMT/MERGA conferences, pp. 760–767). Adelaide, Australia: AAMT and MERGA.