CAL07042

Overcoming Research Design Issues using Rasch Measurement: The *StatSmart* Project

Rosemary Callingham
University of New England
<Rosemary.Callingham@une.edu.au>

Jane M. Watson
University of Tasmania
<Jane.Watson@utas.edu.au>

Longitudinal research in educational settings is notoriously difficult, and when the study is on a large scale, becomes even more problematic. The *StatSmart* project aims to measure changes in both teachers and their students over a period of three years. Using previously validated instruments for students, three overlapping test forms will be used with a rotating design so that each student will only complete each test once. Teacher change will be measured using a profiling instrument at the start, middle and end of the study. Rasch measurement provides for this type of design and should provide good information even with the complexity of students changing classes or schools. The issues associated with complex designs of this nature are discussed.

Criticisms of educational research in recent years have led to calls for improved relevance to the education community and other interested parties. Over ten years ago, Hargreaves (1996) called for educational research to become more relevant to the teaching profession. This call was echoed by Tooley (1998), who suggested that educational research was unfocussed and lacked rigour. As a result of Tooley's critique, research funding policies in Britain were changed to favour studies more applicable to educational contexts.

Encouraging applied studies, however, does not automatically guarantee that the findings will be used. Recently in Australia, Hempenstall (2007) stated

> Education has a history of regularly adopting new ideas, but it has done so without the wide-scale assessment and scientific research that is necessary to distinguish effective from ineffective reforms. This absence of a scientific perspective has precluded systematic improvement in the education system, and it has impeded growth in the teaching profession for a long time. (p.1)

The notion of "scientific" research in education has not been accepted unchallenged. Ball (2001), for example, argued that to discuss educational research in terms only of methodology was to divorce educational research from related fields in the social sciences, and indicated that a range of solutions should be canvassed to maximise the effectiveness of research outcomes to educational problems.

Burkhardt and Schoenfeld (2003) argued that educational research needed to take an "engineering" approach, in which research is " … directly concerned with practical impact—understanding how the world works and helping it 'to work better' by designing and systematically developing high-quality solutions to practical problems" (p. 5). More recently, Baker (2007) in her Presidential address to the American Educational Research Association argued that quality control and feedback loops were needed in research studies to ensure their utility.

Arguing for a scientific approach to educational research can ignore the practical difficulties associated with such designs. The context of education is such that no two classrooms are alike, even when taught by the same teacher in the same way. Comparative studies can never control for every variable, as is the case in scientific experimental studies. Teaching intervention studies, therefore, are difficult to design because so many factors other than the intervention itself may intervene (Shayer, 1992).

Longitudinal studies are also subject to many problems. Tracking students over time, especially across transitions such as that from primary to high school, has a range of issues because of the mobility of students and the fact that classes of students do not stay intact from year to year. Successful longitudinal studies are those that track individuals, using techniques such as telephone or face-to-face interviews (e.g., the Longitudinal Study of Australian Youth, undertaken by the Australian Council for Educational Research, see: http://www.acer.edu.au/lsay/research.html). In addition to the tracking problems, obtaining sound measures of students' achievement over time is also problematic. If the same test is administered results can be contaminated by students' familiarity with items. Different tests, or school achievement results, are usually not on the same scale so that direct comparison across time lacks dependability.

Studies of teacher effectiveness are also difficult. Leigh (2007) recently conducted a study using archived statewide test data from Queensland in which students' test score gains were used as a measure of teachers' effectiveness. The study concluded that very little of the observed variance between teachers was attributable to characteristics available from pay roll databases, such as qualifications or years of service. In particular, qualifications did not appear to affect students' outcomes. Given that it is well established that individual teachers make a difference (e.g., Hill, Rowe, Holmes-Smith, & Russell, 1996), it would seem desirable to try to identify factors that contribute to changes in students' learning outcomes beyond those maintained in employers records. This is one aim of the *StatSmart* project.

Watson, Beswick, Brown, & Callingham (in press) attempted to consider student changes in attitudes and achievement against a general background of teachers' professional development. The small scale study was limited in the inferences that could be drawn, in particular by the inability to link teachers directly to students because of the nature of the study. One aim of *StatSmart* is to address this issue in order to examine some of the factors in teacher behaviour that affect students' outcomes.

The *StatSmart* project

StatSmart is a three-year, ARC Linkage funded project involving 17 primary and high schools in three states, South Australia, Tasmania and Victoria. The major educational aim of this research is to improve the statistical understanding of school students through the improvement of the teaching of statistics at the middle and high school levels. The project will document and evaluate teachers' professional learning in terms of the connections required among the seven forms of knowledge suggested by Shulman (1987a, 1987b): content knowledge, general pedagogical knowledge, curriculum knowledge, pedagogical content knowledge, knowledge of learners and their characteristics, knowledge of education contexts, and knowledge of education ends, purposes, and values.

Achieving this outcome, however, is not a trivial task. The *StatSmart* project is taking a Teacher-Researcher approach in which the 50 or so teachers involved will identify specific issues within their own contexts and develop strategies to address these, using a broad theoretical framework of hierarchical understanding of statistical concepts developed from previous studies (e.g., Callingham & Watson, 2005; Watson & Callingham, 2003; Watson, 2006). This approach has been shown to be successful in linking theory with practice in recent design studies (e.g., Shavelson, Phillips, Towne, & Feuer, 2003) and to bring about changes in teachers' behaviour that lead to improved learning outcomes (Hiebert, 1999). To support their efforts in schools, teachers will receive state-of-the-art educational software, *TinkerPlots* (Konold & Miller, 2005), use real data sets from the *CensusAtSchool* website (http://www.abs.gov.au/websitedbs/cashome.nsf/Home/Home), be given recent research information about students' development of statistical understanding (Watson, 2006) and attend a two-day conference each year at which leaders in the field of statistic education in schools will provide workshops and teaching ideas. Teachers will make operational the ideas and skills that they learn in their own contexts and this will inevitably lead to considerable variation in outcomes, both for teachers and students. The challenge for the researchers is to measure these outcomes longitudinally in such a way that meaningful inferences can be drawn about models of professional learning, and aspects of teachers' knowledge that have an impact on students' outcomes.

The scale of the project, three states, 17 schools and around 50 teachers, adds to the complexity. Each state has a different curriculum framework within which statistics is taught. Primary and high schools from state, Catholic and independent sectors are involved, and the teachers have different levels of experience, backgrounds and training. The intention of the project coordinators is to undertake research in "real world" settings and "scale up" previous studies in order to identify aspects of professional learning that are amenable to systemic intervention to bring about changes in students' learning.

With respect to professional learning in the context of mathematics curriculum change in the United States, Mewborn (2003) claimed that descriptions of professional development programs for mathematics teachers abound; however, she did not find any studies claiming long-term change for teachers or students, using any criteria. Other studies of professional learning for mathematics and science teachers found teacher change to be only short term (Loucks-Horsley & Matsumoto, 1999) or found outcomes based only on opinions and participation (Schoen, Cebulla, Finn, & Fi, 2003). The need for teachers to have appropriate content and pedagogical knowledge is emphasised in standards for teacher competency (e.g., Australian Association of Mathematics Teachers, 2000; New South Wales Institute of Teachers, 2003) yet there are few studies explicitly linking teachers' improved knowledge to students' learning outcomes. These observations are confirmed in the draft report of the Australian Councils of Deans of Education and Deans of Science (2003) on professional learning for mathematics, science, and technology teachers, which called for hard evidence of long-term teacher and student change. It is this challenge of providing evidence of teacher and student change that *StatSmart* aims to address.

The *StatSmart* Research Design

To meet the demands for an engineering approach, *StatSmart* uses a design experiment methodology (Brown, 1992). Design experiments are described as

attempts "to engineer innovative educational environments and simultaneously conduct experimental studies of those innovations" (p. 141). Such an approach requires accepting a complex educational setting as a holistic system, in which it is not possible to change one component without concomitant changes elsewhere. The need is to collect data that will allow inferences to be drawn about these effects that can contribute to theoretical positions as well as providing practical solutions. The demands of such a design are immense, and provide considerable challenges for the researchers. In *StatSmart*, the hope is that the variety and nature of the data collected will provide appropriate information about the synergistic effects of teacher change, technological approaches to learning statistics and use of real world data in classrooms.

Instruments

Teacher and student change will be tracked over time using previously validated instruments. Teachers will complete a profiling instrument (Watson, 2001; Watson, Beswick, Caney, & Skalicky, 2006) that addresses Shulman's (1987a, 9187b) seven knowledge components. This instrument consists of survey and interview components. The survey part includes sections using Likert scales that address confidence in teaching different statistical topics and attitudes and beliefs about statistics. Two sections directly address teachers' own content knowledge and content pedagogical knowledge. In the first, items from media surveys are presented and teachers are asked to indicate how students might respond both appropriately and inappropriately to them. Teachers not only must recognise the statistical knowledge addressed but also be able to identify likely student development of understanding. The second section presents two real responses from students to items addressing aspects of chance and data in the classroom. Teachers are asked to indicate what they would do to intervene to improve students' understanding. General pedagogical knowledge is covered by two sets of items asking about teaching and assessment techniques. Other aspects of teacher knowledge, such as curriculum knowledge and understanding of educational contexts are covered in the interview component of the profile.

Students will complete surveys three times during the project. The survey instruments have been used previously to establish a development scale of statistical literacy (Callingham & Watson, 2005; Watson & Callingham, 2003, 2005; Watson, Kelly, Callingham, & Shaugnhessy, 2003). They cover six components of statistical understanding, identified by Holmes (1980): data collection, data tabulation, data representation, data reduction, probability, and interpretation and inference. All questions will be scored using rubrics developed in previous projects that have been shown to provide reliable data when analysed using the Partial Credit Model (Masters, 1982).

Student Survey Administration

It is often difficult to track students over time. They change schools, are absent when the survey is administered, and disperse into different schools at the transition point from primary into secondary schooling. In an attempt to overcome some of these problems, baseline data will be collected as early as possible in the year in which they enter the *StatSmart* project, and a second collection of data, for monitoring purposes, will occur as late as possible in the same year. It is hoped that this will capture most students with at least two surveys. A longitudinal survey will be given to

as many students as possible after 12 months. This design, however, carries the dangers of students' recognising and remembering the items if the same survey form is used each time. To overcome this problem, Rasch measurement equating techniques using common item linking will be used (Kolen, 1999).

The items have been organised into three test forms of between 24 and 29 items, of which 12 are link items. The link items were chosen because they showed good fit and a range of difficulties in previous studies. All test forms provide a balance of items across Holmes' (1980) categories. In this way, survey formats have been constructed to address the needs of the *StatSmart* study, and to conform to the requirements of Rasch measurement. To reduce further the possibility of students recognising items, the link items have been dispersed in different places in each test form. A rotating test design is used so that every student should answer each test form once only. There is added complexity in that teachers in the project are likely to have new classes each year, and the possibility of tracking a single class of students across three years is limited. It is probable that new students, and possibly teachers, will enter the project each year. Hence, baseline data will be collected from these new students at the start of each year, and the monitoring test for these students will also provide the longitudinal data for those students already in the project. A summary of this design is shown in Figure 1, showing the purpose of the test and the rotation across the three states involved.

Year	SA	VIC	TAS	Purpose	
Mar-07	Test A	Test B	Test C	Baseline data	
Nov-07	Test C	Test A	Test B	Monitor	
Mar-08	Test A	Test B	Test C	Baseline data	NB Taken only by students new to *StatSmart*
Nov-08	Test B	Test C	Test A	Longitudinal	Students in project in 2007
	Test B	Test C	Test A	Monitor	Students new in project in 2008
Mar-09	Test A	Test B	Test C	Baseline data	NB Taken only by students new to *StatSmart*
Nov-09	Test C	Test A	Test B	Longitudinal	Students in project in 2008
	Test C	Test A	Test B	Monitor	Students new in project in 2009

Figure 1. Summary of the rotating test design for *StatSmart*.

Teacher Profile Administration

As indicated earlier, teachers will also be surveyed three times during the *StatSmart* study, once in each year. Teachers who join the project after the first year will complete as many Teacher Profiles as their involvement allows. With over 50 teachers participating in Year 1 of the research, it is hoped that a large enough number will be retained through the full course of the project to enable meaningful inferences about teacher change to be drawn from the data collected. It is not proposed to change the teacher profile instrument during the course of the study. Unlike the student surveys, the range of question types and content addressed make it less likely that teachers will remember their specific responses from year to year. If this does become an issue, however, the possibility is available for new items to be introduced, especially in the pedagogical content knowledge aspects of the profile.

There are a number of specific aspects that are of interest arising from the Teacher Profile. Shulman's (1987a, 1987b) components of teacher knowledge are sufficiently different to raise issues of dimensionality. A combination of Rasch methodology and Principal Component Factor Analysis (Grimbeek & Nesbit, 2006) will allow a

consideration of the cohesiveness or otherwise of Shulman's categories. This may have relevance to the current discussion about teacher standards.

Os specific interest to the *StatSmart* project, however, is the extent to which teacher change brings about improved student outcomes. Teachers will be linked to particular classes and hence to individual students in each year of the project. In this way it is hoped to identify both effective teaching practice and aspects of the professional development program that have an impact on students' outcomes. The structure of the Teacher Profile instrument also makes it possible to identify which aspects of Shulman's knowledge categories are key to bringing about changes in students.

Qualitative Data

In addition to the formal measurement instruments, interviews with students and classroom observations will provide information about the nature of the teaching and learning experiences of students. Interviews will focus on uncovering students' understanding of statistical ideas, using existing protocols (Watson, Callingham, & Kelly, 2007; Watson & Moritz, 2000, 2003), and some new questions that will focus on the use of technology in learning statistics. Students' work samples, examples of teachers' planning and assessment, and curriculum documents will also be collected to provide explanations of changes that are observed from the teachers' profiles and students' surveys. These qualitative data will add rich detail to the quantitative data, and provide opportunities to explain findings.

Use of Rasch Measurement

The *StatSmart* research design is underpinned by the use of Rasch measurement (Bond & Fox, 2007). The Rasch model (Rasch, 1960) uses the interaction between students (or test takers) and items to create an interval level scale. The original dichotomous model has been extended to include the Partial Credit Model (PCM) in which the step structure of the items can vary from item to item (Masters, 1982).

The PCM can be expressed as:

$$\frac{\pi_{ix}}{\pi_{i(x-1)} + \pi_{ix}} = \frac{\exp(\beta - \delta_{ix})}{1 + \exp(\beta - \delta_{ix})}$$

where

π_{ix} is the probability of a person responding in category x (x = 1,2, … m) of item i;

β is the person's ability in the domain being measured by this set of items; and

δ_{ix} is the difficulty of the step threshold that governs the probability of the response occurring in category x rather than category x–1.

In the PCM, the probability of a student responding in the x^{th} category, as opposed to the $x\text{-}1^{th}$ category, is dependent on the difficulty of the x^{th} level. By using an estimate of "step difficulty" within each item in the assessment, the PCM locates a person on an underlying variable through a consideration of the number of steps that the person has made beyond the lowest level of performance. Using assessment items that address increasing competence on an underlying variable, or latent trait, that are scored with more than a right/wrong response, the model provides information about a student's level of understanding against the target variable. The points at which the

likelihood of a higher-level response became greater than that of a lower-level response are called thresholds. In the context of a particular activity, the threshold is the point on the variable at which the probability of being observed in or above a particular score code category changes from less than 50% to 50% and above (Wright & Masters, 1982). A students' test score (ability measure) is scaled in logits, the logarithm of the odds of success and the unit of Rasch measurement, and in an analogous operation item difficulty can be expressed similarly.

Numerous studies have used Rasch measurement approaches to consider longitudinal change (e.g., Callingham & Griffin, 2002; Griffin & Callingham, 2006; Watson, Kelly, & Izard, 2006) The use of Rasch measurement has a number of benefits. The creation of a "ruler" and the capacity to anchor item difficulties in order to calibrate to the same scale means that changes in performance can be compared with confidence across time without having to use the same test (Bond & Fox, 2007, ch. 5). In addition, the measures of students' abilities can be used in a variety of ways. Apart from the usual statistical comparisons of means, "gain scores" can be obtained from the change in the logit measure between administrations. These may be used to determine "value-added" measures, such as those described by Leigh (2007).

Of more use to teachers, however, is the creation of a "profile" of student achievement (Griffin & Jones, 1987). The underlying variable can be "segmented" (Wilson, 1999) through a process of analysing the clusters of items along the variable produced by Rasch modelling. By asking questions such as "What do these items have in common?" and "How do the skills and knowledge demands of this cluster differ from those of that cluster?", a set of criteria can be established that define the qualitative changes in the demands of the items, and thus describe levels of response. This approach is in line with Fisher's (1994) discussion of the qualitative aspects of the Rasch model and provides a conceptual interpretation of the variable that can be compared with the theorised construct. This profiling approach was used in previous studies using the *StatSmart* items (Callingham & Watson, 2005; Watson & Callingham, 2003; Watson & Callingham, 2005). The advantage to teachers is that "nutshell" statements (Griffin, Smith & Ridge, 2001) about their students' achievements provide a useful indication of students' progress without swamping them with too much detail.

In addition, as well as differences in mean scores, the proportion of students in the levels or bands of the profile can provide a different indicator of change over time. Rather than using means, which are subject to outliers, or percentiles, which provide a statistical approach to grouping students' scores, considering groupings of students against profile bands or levels links students' achievements directly to a conceptual interpretation of the underlying variable. This approach provides powerful information to teachers and systems, and has been used in a number of studies (Callingham & McIntosh, 2002; Griffin, 2001).

Another benefit of using Rasch measurement is the use of differential item functioning (DIF) to identify bias. If the probabilities of response to an item cannot be explained wholly by the ability of the student and the fixed difficulty parameters of the item, the item is considered to exhibit DIF (Wu, Adams & Wilson, 1998). DIF analysis provides a comparison of the behaviour of items with respect to defined groupings of students. It can reveal systematic differences in the ways in which these groupings interact with the items. In *StatSmart*, this may be of particular interest across states because of the different curriculum frameworks that are used.

Data Issues Arising from the Design

The complex nature of the *StatSmart* design raises a number of technical issues. The sample of students and teachers is out of the control of the researchers, and may vary from year to year. Relying on a convenience sample inherently introduces potential bias. On the one hand, teachers and hence students are in schools where there must be some systemic support from the school executive. This support is more likely to create a situation in which the project can succeed. On the other hand, the expectation that at least two and preferably three teachers and their classes will be involved inevitably has led to some teachers being reluctant participants. This situation, however, is the same as that faced by any system implementing a professional development program.

Linking teachers to students in classes, and tracking both students and teachers over time will provide a challenge. Carefully designed identification numbers for teachers and students will be used and all records will be stored in a relational database. Nevertheless, this aspect will be one of the most challenging of the project.

Such a design would not be possible without the use of Rasch measurement. Traditional longitudinal research designs have suffered from having to use the same items in order to monitor change. Using identical test forms each time can lead to familiarity, and restricts the inferences that can be drawn.

There is a further advantage to the use of Rasch measurement. All the items have been used in several previous studies involving large numbers of students. It is possible to use anchor values from these earlier studies to provide information about the students in *StatSmart* in relation to students in these earlier projects. Although this does not provide a genuine comparative group, it goes some way towards meeting objections about the shortage of comparative studies. The anchoring process also provides for genuine comparisons of students across time. In addition, the flexibility that Rasch measurement affords of using different types of data, such as student interview and survey data, which are all underpinned by developmental models, is another strength of the Rasch approach.

Design research requires good quality data that can be used in multiple ways such as reporting back to teachers or stretching the boundaries of theoretical knowledge (Brown, 1992). Rasch measurement provides the necessary rigour but allows for a qualitative interpretation of the variable as well as sound statistical measures. As such it is the ideal vehicle for a challenging, large scale project such as *StatSmart*.

Acknowledgement

The *StatSmart* study is funded by Australian Research Council grant number LP 0669106.

References

Australian Association of Mathematics Teachers, Inc. (2000). Consultation draft descriptors of excellence in teaching mathematics. Adelaide, SA: Author.

Australian Council of Deans of Education & Australian Council of Deans of Science. (2003). Professional learning for enhancing teaching and learning within science, mathematics and technology in Australia. [Draft report] Canberra: Author.

Baker, E. (2007). *The end(s) of testing. Presidential address to the Annual Meeting of the American Educational Research Association, Chicago*. Available at http://www.softconference.com/Media/WMP/270409/s40.htm

Ball, S. J. (2001). You've been NERFed!' Dumbing down the academy: National Educational Research Forum: 'a national strategy ? consultation paper': a brief and bilious response. *Journal of Education Policy, 16*(3), 265– 268.

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

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*(2), 141-178.

Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research: Toward a more useful more influential, and better-funded enterprise. *Educational Researcher, 32*(9), 3-14. Retrieved 4 February, 2004 from http:// www. aera.net/pubs/er/eronline.htm

Callingham, R. & Griffin, P. (2002). Summary report of the INISSS project, 1999-2001. Armidale: UNE. Accessed 2nd March 2005 from http://www.blackdouglas.com.au/taskcentre/tdocs.htm

Callingham, R. & McIntosh, A. (2002). Mental computation competence across years 3 to 10. In B. Barton, K.C. Irwin, M. Pfannkuch & M.O.J. Thomas, (Eds.), *Mathematics education in the South Pacific.* (Proceedings of the 25th Annual Conference of the Mathematics Education Research Group of Australasia, pp. 155-163). Sydney: MERGA.

Callingham, R. & Watson, J. M. (2005). Measuring statistical literacy. *Journal of Applied Measurement, 6 (1)*, 29, 19-47.

Fisher, W.P. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.) *Objective measurement Vol. 2* (pp. 36 – 72). Norwood, NJ: Ablex.

Griffin, P. (2001). Levels of competence in literacy and numeracy for SACMEQ II. Paper presented at the Working Meeting of SACMEQ National Research Co-ordinators on "SACMEQ II Policy Report Preparation". Paris: France, 10-15 October.

Griffin, P. & Callingham, R. (2006). A twenty-year study of mathematics achievement. *Journal for Research in Mathematics Education, 37* (3), 167-186.

Griffin, P. & Jones, C. (1987, November). *Assessing the development of reading behaviours: A report of profiles and reading band development.* Paper presented at the annual conference of the Australian Association for Research in Education. University of New England, Armidale, NSW.

Griffin, P., Smith, P., & Ridge, N. (2002). *Literacy profiles in practice*. Portsmouth, New Hampshire: Heinemann.

Grimbeek P., & Nesbit, S. (2006). Surveying primary teachers about compulsory numeracy testing: Combining factor analysis with Rasch analysis. *Mathematics Education Research Journal*, *18*(2), 27-39 .

Hargreaves, D. (1996) *Teaching as a research-based profession: possibilities and prospects*. (Teacher Training Agency Annual Lecture). London: Teacher Training Agency.

Hempenstall, K. (2007). *Submission to the Senate Committee on Academic Standards in School Education*. Available at http://www.aph.gov.au/SEnate/committee/eet_ctte/academic_standards/submissions/sublist.htm

Hiebert, J. (1999). Relationships between research and the NCTM Standards. *Journal for Research in Mathematics Education, 30*, 3-19.

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* (Vol. 1). Melbourne: Centre for Applied Educational Research, University of Melbourne.

Holmes, P. (1980). *Teaching statistics 11-16*. Slough, UK: Schools Council and Foulsham Educational.

Kolen, M. J. (1999). Equating of tests. In G.N. Masters & J.P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 164 – 175). New York: Pergamon.

Konold, C., & Miller, C.D. (2005). *TinkerPlots: Dynamic data exploration*. [Computer software] Emeryville, CA: Key Curriculum Press.

Leigh, A. (2007). Estimating teacher effectiveness from two-year changes in students' test scores. Available at: http://econrsss.anu.edu.au/~aleigh/pdf/TQPanel.pdf

Loucks-Horsley, S., & Matsumoto, C. (1999). Research on professional development for teachers of mathematics and science: The state of the scene. School Science and Mathematics, 99, 258-271.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 49*, 359-381.

Mewborn, D.S. (2003). Teaching, teachers' knowledge, and their professional development. In J. Kilpatrick, W.G. Martin, & D. Schifter (Eds.), A research companion to Principles and Standards for School Mathematics. Reston, VA: National Council of Teachers of Mathematics.

New South Wales Institute of Teachers. (2003). Draft Teaching Standards Framework – 30 June 2003. Sydney: Author. Retrieved 16 February, 2004 from http://www.icit.nse.edu.au

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Schoen, H.L., Cebulla, K.J., Finn, K.F., & Fi, C. (2003). Teacher variables that relate to student achievement when using a standards-based curriculum. Journal for Research in Mathematics Education, 34, 228-259.

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of education design studies. *Educational Researcher, 32*(1), 25-28.

Shayer, M. (1992). Problems and issues in intervention studies. In A. Demetriou, M. Shayer,. & A. Efklides, (eds.) *Neo-Piagetian theories of cognitive development: implications and applications for education* (pp.107-121). London: Routledge.

Shulman, L. S. (1987a). Assessing for teaching: An initiative for the profession. *Phi Delta Kappan, 69*(1), 38-44.

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

Tooley, J. (1998). *Educational research: A review*. London: Office for Standards in Education and HMSO.

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: Lawrence Erlbaum.

Watson, J., Beswick, K., Caney, A., & Skalicky, J. (2006). Profiling teacher change resulting from a professional learning program. *Mathematics Teacher Education and Development, 7*, 3-17.

Watson, J., Kelly, B., & Izard, J. (2006). A longitudinal study of student understanding of chance and data. *Mathematics Education Research Journal*, *18*(2), 40-55.

Watson, J.M., & Callingham, R.A. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3-46.

Watson, J.M., & Callingham, R.A. (2005). Statistical literacy: From idiosyncratic to critical thinking. In G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education. International Association for Statistical Education (IASE) Roundtable, Lund, Sweden, 2004* (pp. 116-162). Voorburg, The Netherlands: International Statistical Institute.

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

Watson, J.M., & Moritz, J.B. (2003). Fairness of dice: A longitudinal study of students' beliefs and strategies for making judgments. *Journal for Research in Mathematics Education*, *34*, 270-304.

Watson, J.M., Beswick, K., Brown, N., & Callingham, R. (in press). Student change associated with teachers' professional learning. *Proceedings of the 30th annual conference of the Mathematics Education Research Group of Australasia.*

Watson, J.M., Callingham, R. & Kelly, B.A. (2007). Students' appreciation of variation as a foundation for statistical understanding. *Mathematical Thinking and Learning, 9*(2), 83-130.

Watson, J.M., Kelly, B.A., Callingham, R.A., & Shaughnessy, J.M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology*, *34*, 1-29.

Wilson, M. (1999). Measurement of developmental levels. In G.N. Masters, & J.P. Keeves (Eds.). *Advances in measurement in educational research and assessment.* (pp. 151-163). New York: Pergamon.

Wright, B. & Masters, G. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wu, M.J., Adams, R.J. & Wilson, M.R. (1998). *ACER Conquest. Generalised item response modelling software manual*. Melbourne: ACER Press.