The Development and Validation of the Students' Self-efficacy for Statistical Literacy Scale

Colin Carmichael
University of Tasmania
<colin.carmichael@utas.edu.au>

Ian Hay
University of Tasmania
<Ian.Hay@utas.edu.au>

Statistical literacy is a comparatively new concept in mathematics and while there is some consensus about how it is defined, there has been limited research on how the concept is measured within a school context. This paper, reports on the development and validation of an instrument to measure middle school students' self-efficacy for statistical literacy. The items were developed from the relevant research literature and then tested on a sample of 366 students. A Rasch measurement methodology was used to create the measure and to provide evidence for its construct validity. The evidence reported in this paper indicates that the proposed instrument has suitable reliability and validity properties.

The ability to interpret and critically evaluate messages that contain statistical elements, termed *statistical literacy* (Gal, 2003), is paramount in our information rich society. The foundation knowledge and skills of this literacy are embedded in the chance and data strand of current Australian mathematics curriculum. The importance of this literacy is reflected in the fact that chance and data is one of only three content strands in the proposed national mathematics curriculum (National Curriculum Board, 2008). While researchers have investigated student cognitive development in statistical literacy (Callingham & Watson, 2005; Watson, 2006), few have explicitly investigated the associated affective development and its dimensions. Yet affect plays an important part in students' learning, with Panksepp (2003) arguing that many higher order cognitive abilities co-evolve with corresponding affective processes. Further, the indications are that early adolescence appears to be a critical stage in the affective development of students (Watt, 2008), with the correlation between students' attitudes towards mathematics and their achievement in mathematics, strongest for students in this developmental period (Ma & Kishor, 1997). This paper reports one aspect of a larger study that seeks to investigate the development of middle school students' interest in statistical literacy. As is detailed in Carmichael and Hay (2008), such interest is regarded as an affect.

The suggestion is that students' beliefs regarding their competency in secondary mathematics mediate the relationship between their interest, knowledge and achievement (Trautwein, Ludtke, Köller, Marsh, & Baumert, 2006). Students are motivated to engage in tasks that they find interesting. This may not occur, however, if the student believes success is unlikely. Studies that have investigated this relationship typically examine students' academic self-concept, which is often assessed using broad items, such as "mathematics is my best subject". Ma and Kishor (1997) argued that such broad items provide only crude approximations and recommended items should instead target the specific topics and activities that comprise mathematics learning. In addition to this, academic self-concept is a past orientated construct (Bong & Skaalvik, 2003) with students reflecting on their past experiences rather than considering future achievement. In this study, students' beliefs regarding their competency are assessed through a construct termed *self-efficacy*, which is defined as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura, 1997, p.3). This construct is future orientated (Bong & Skaalvik, 2003) and is typically assessed through items that ask students to indicate their level of confidence in achieving specific rather than

In R. Hunter, B. Bicknell, & T. Burgess (Eds.), *Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 1). Palmerston North, NZ: MERGA.

general tasks. Of all the psychosocial factors, self-efficacy is considered to be one of the best predictors of achievement in an educational context (Robbins et al., 2004). Consequently it is likely that self-efficacy will provide more insight into middle school students' interest development than other measures of students' self-competency beliefs.

Although instruments have been developed to assess self-efficacy in a number of contexts, none noted have specifically been developed for a statistical literacy context. Accordingly the aim of this study is to develop a self-efficacy subscale for a larger interest questionnaire that will be used in a middle school statistical literacy context. This paper reports on the development of this subscale, termed the Self-efficacy for Statistical Literacy (SESL) subscale.

Methodology

Development of Items

It is expected that students' self-efficacy towards statistical literacy will be influenced by the specific topics that comprise statistical literacy. These topics are identified by Watson (2006) as: sampling or data collection; graphs or data presentation; average; chance; beginning inference; and, variation. Given the age of students in this study, however, many would not be sufficiently cognizant with the term variation to respond meaningfully to self-descriptions from that topic. Accordingly the topics in the current study are restricted to the first five of Watson's topics, reflecting an earlier classification by Holmes (1986). Tasks from some of these topics are typically more difficult than those from others. For example, in the development of their statistical literacy hierarchy, Callingham and Watson (2005) observed that average problems are first successfully encountered by students near the middle of the hierarchy. In contrast, students at the very lowest levels of the hierarchy can successfully read a number from a table of numbers. It is expected that students' self-efficacy towards statistical literacy should reflect these differences.

Context also plays a particularly key role in the development of statistical literacy. Watson (2006) argued that students at higher levels of the statistical literacy hierarchy are more able to interact critically with the contexts in which tasks are situated. Students' self-efficacy towards statistical literacy, therefore, should be influenced by the context in which the task is situated. Contexts, however, are chosen by teachers to suit the specific needs of their students and can vary widely. As a result, this study has focussed on more general contexts, in particular those that are school-, and media-related.

A bank of twenty items was written to reflect the above topics and contexts used in the teaching of statistical literacy. This was then subject to a panelling process whereby the appropriateness of items was reviewed by a number of academics and practicing teachers. At this stage it was decided that a greater proportion of items assess student self-efficacy towards graphs, as this reflects an emphasis on this aspect of statistical literacy in the middle school context. After the review a sample of nine items was used for piloting. These were written in the form of self-descriptions with the common stem "I am confident that I am able to". Students were required to gauge how closely they could identify with each self-description using a five-point Likert scale ranging from 1 ("Not me at all") to 5 ("Describes me well").

Piloting of Instrument

Piloting occurred in two phases over a period of eight months. During the first phase, which was based in Queensland, item development continued. This was directed by teacher feedback and the results of statistical analysis. In the second phase, however, the finalised instrument was tested on a group of students from other Australian states. A total of 711 students from 11 schools across four Australian states, were invited to participate in the study. The results reported here are based on 366 complete responses, a response rate of 51%. Of this sample: 44% were male; ages ranged from 11.3 to 16.0 years with an average of 13.6; 29% of students came from single sex schools and 71% from co-educational schools; 43% of students came from government schools and 57% from independent schools. Specific details for each phase of the study are shown in Table 1.

Table 1
Student and school details for each phase of the study

Phase	Student details			School details	
	Number	Mean age (yrs)	Males (%)	Number	States involved
1	221	13.3	35	6	QLD
2	145	13.9	54	5	SA, VIC, TAS

Students in the first phase were also required to respond to a sample of items from the previously validated "Motivated Strategies for Learning Questionnaire" (Pintrich & De Groot, 1990) in order to provide a measure of concurrent validity. The self-efficacy subscale of the Motivated Strategies for Learning Questionnaire (MSLQ) provides a measure of general self-efficacy towards mathematics. Given that Australian students currently learn most of their statistical concepts during mathematics lessons, it is argued that their mathematics self-efficacy should correlate positively with their self-efficacy for statistical literacy.

Analysis of Student Responses

Student responses were analysed using the *Rasch Rating Scale Model* (Andrich, 1978). This statistical model seeks to estimate a number of parameters from the data that include: the self-efficacy level of the students, the perceived difficulty of each item, and the relative degree of difficulty between successive Likert responses. Given that the model fits the data, it will produce an interval measure of self-efficacy: A scale upon which both student self-efficacy and the perceived difficulty of items can be placed. Model fit is assessed through an analysis of the residuals, where each residual is the difference between an observed and an expected outcome. Fit statistics that are based on these residuals are then calculated. This study reports the *infit* statistic (denoted v), which is an information (or variance) weighted sum of squared standardized residuals. For rating scales, such as the one used in this study, Bond and Fox (2007) recommended that $0.6 \leq v \leq 1.4$.

Student responses for all items in this study were analysed using the Rasch modelling program *Winsteps* (Linacre, 2006). In addition to the model fit statistics described above, this program also provides an estimate for Cronbach's coefficient of reliability (α) and a number of statistics that include the *strata*, which is defined as the number of distinct levels of person ability that the items are able to distinguish (Smith, 2001).

During the analysis stage of phase 1 and based on reported fit statistics, items with poor fit were routinely analysed and in some cases removed. Such an approach ensures that the observed data are as close as possible to a *conjoint system* (Luce & Tukey, 1964), one in which the variables can be considered to be quantitative and thus measurable on an interval scale. This process, however, is balanced by the need to maintain content validity.

Results

The Development of Items During Phase 1

The 9 items comprising the SESL, together with their topic and context, are shown in Table 2. Three of these items, however, were changed during the first phase of the pilot on the basis of 140 complete student responses. Initially item 1 was worded "I am confident that I can correctly calculate the average of 8 exam results". This item displayed significant misfit ($v = 1.6$) and was subsequently altered to the current wording. The intent was to produce an item of greater difficulty and better fit. Similarly, item 7 initially assessed students' confidence to explain the term "random", as knowledge of such terms is regarded as fundamental to statistical literacy (Watson, 2006). This item, however, also displayed significant misfit ($v = 1.42$). It was decided to instead include an item that assessed beginning inference, which is currently shown as item 7. In addition to these two changes, it was felt that the measure needed an easier item. At that stage the reported strata statistic was less than 3, suggesting the need for both easier and more difficult items (Smith, 2001). Item 8, which originally assessed confidence in calculating probabilities associated with dice or coins, was duplicated to an extent in item 3. This item was removed and replaced by the current item 8, which assesses confidence in arranging data into tables, a task that students near the lower reaches of the statistical literacy hierarchy should readily achieve (Callingham & Watson, 2005).

Table 2
Items comprising the SESL

No.	Item ("I am confident that I am able to:")	Topic/context
1	Solve problems that use averages	Average/school
2	Find when a newspaper article has used the wrong type of average.	Average/media
3	Explain to a friend how probability (or chance) is calculated.	Chance/school
4	Show data correctly on a bar chart.	Graphs/school
5	Explain the meaning of a graph in a newspaper or on the internet.	Graphs/media
6	Find a mistake in someone else's graph.	Graphs/both
7	Explain when conclusions that are based on surveys might be wrong.	Inference/both
8	Arrange my data correctly into a table.	Graphs/school
9	Explain how to select a fair sample of students for a school survey.	Sampling/school

The Measure and its Statistics

Having finalised the items that comprise the SESL during phase 1, the instrument was subsequently tested on an independent group of students during phase 2. Statistical analysis was then conducted on the data from both phases separately and also the pooled

results from both phases. Based on the pooled results, the instrument provides a measure that explains 70% of the variance in student responses with a reported reliability coefficient of $\alpha = 0.93$. Estimates of item difficulties are shown in Table 3 for both phases and the pooled results. The reported standard errors (SE), and item fit statistics are based on the pooled results. The difficulty estimates for the measure, as calibrated from the pooled sample, range from -0.61 to 0.75 logits. The fit statistics lie within acceptable limits and indicate that the measure approximates a conjoint system.

Although the apparent range of item difficulties is only 1.36 logits, these reported item difficulties are actually mid-points of the difficulty levels that are associated with each of the five Likert responses to the item. Consequently the instrument is actually able to detect a much larger range of self-efficacy. In fact the strata statistic is reported as 3.5, which indicates that the measure can differentiate between 3 and 4 statistically distinct student self-efficacy levels (Smith, 2001).

The model was applied to different subsets of students in order to assess whether items functioned differently for such groups. In particular item difficulties were estimated for males and females, and differences in these estimates were assessed for statistical significance. There was no evidence for gender differences in the items. There was, however, evidence that younger students found item 8 more difficult than older students. For example, the estimated difficulty of this item for year 9 students in phase 1 was -0.9, while for year 7 students it was -0.45.

Table 3
Difficulty estimates and fit statistics for items comprising the SESL

Item	Estimated item difficulty			SE	No. students	Infit v
	Phase 1	Phase 2	Pooled			
1	-0.49	-0.71	-0.58	0.08	221	1.16
2	0.8	0.66	0.75	0.06	362	0.93
3	0.28	0.33	0.29	0.06	362	1.01
4	-0.56	-0.53	-0.56	0.06	362	1.15
5	0.19	0.14	0.16	0.06	361	0.89
6	0.00	0.05	0.01	0.06	361	1.09
7	0.63	0.54	0.54	0.08	222	0.81
8	-0.86	-0.49	-0.61	0.08	222	0.77
9	0.01	-0.01	-0.01	0.06	362	1.06

A factor analysis of the residuals indicated that of the 30% of variance that was unexplained, the largest retrieved factor only contributed 5% towards the total variance. Such a level is regarded as indistinguishable from general statistical noise (Linacre, 2006). Consequently this evidence supports the assumption that the measure consists of one underlying dimension.

The rating scale model was also applied to student responses to items in the MSLQ and a mathematics self-efficacy measure was extracted. This measure explained 78% of the variance in student responses and reported a reliability coefficient of $\alpha = 0.95$. Student mathematics self-efficacy scores were found to be positively associated with their SESL scores ($r = 0.56$, $p = 0.00$).

Discussion

Messick (1995) argued that there are six aspects, or forms of evidence, to a validity argument: content, substantive, structural, generalizability, external and consequential. The following discussion examines each of these aspects in relation to the measure reported in this paper.

Content evidence includes arguments that relate to the relevance, representativeness and technical quality of the items (Messick, 1989). The initial paneling process and subsequent refinement of items contributed to their relevance. As is seen from Table 2, several items reflect the interpretative and evaluative aspects of statistical literacy. For example "finding a mistake" or "explaining why" are both fundamental to this literacy. The items were also representative in that they sampled each of the five identified topics of statistical literacy. In addition to this the items span a range of difficulties across the self-efficacy scale. There is, however, a cluster of items at the lower end of the scale suggesting some redundancy. Items 1, 4, and 8 all measure similarly low levels of self-efficacy; this is not surprising in the case of items 4 and 8 which should be of similar difficulty. The ease at which students reported that they can solve problems involving averages, though, suggests that item 1 needs further refinement. In a Rasch measurement paradigm, evidence to support the technical quality of items is provided in the reported fit statistics (Smith, 2001). The fit statistics reported in Table 3 are all within an acceptable range, thus demonstrating that most students respond to the items in a similar way.

Substantive evidence refers to the extent to which underlying theories predict the observed outcomes (Messick, 1995). In this instance the analysis focused on the agreement between the observed and expected hierarchy of item difficulties. The difficulty hierarchy of SESL agrees, in the main, with the statistical literacy hierarchy, as it is reported in Callingham and Watson (2005). The most difficult items in the SESL were items 2 and 7. It is expected that students would regard the evaluation of an external and presumably reliable source as very difficult. Similarly the ability to find mistakes in conclusions that are based on surveys would also be a difficult task. Students at this age should find tasks related to graphs relatively easy as they encounter such tasks early in the mathematics curriculum. Consequently finding a mistake in someone else's graph (item 6) should be easier than explaining to a friend how to calculate a probability (item 3), which is encountered later in the curriculum. Similarly, it is not surprising that arranging data into a table is the easiest task. It is surprising; however, that solving problems involving averages (item 1) has the same reported difficulty as simple tabulation (item 8).

Structural evidence refers to the extent to which the internal structure of the measure reflects the theoretical structure of the construct (Messick, 1995). The results of the factor analysis of the residuals, reported in the last section, support the uni-dimensionality assumption of the measure, as does the fact that the measure explains 70% of the variation in student responses. Similarly, the reported internal consistency of the items ($\alpha = 0.93$) was at a very high level.

Generalizability evidence refers to the extent to which the findings from this sample of items and students can be applied to the construct in other samples of students. A simple test of the generalizability of the measure is to examine the invariance of item difficulty estimates between two samples of students (Smith, 2001). Based on the item difficulties for each phase, as reported in Table 3, both samples of students responded in a similar way to most items. The estimated difficulties of all items except 1 and 8 differed by no more than 2.5 times the pooled standard error. Students in phase 1 of the study were much more confident at arranging their data into tables than those in the second phase. In contrast,

students in the second phase were much more confident at solving problems that involved averages. This anomaly is partly due to differences in the sample. The proportion of females in the first phase was considerably higher than that in the second phase; they were also younger. Arguably the older students in phase 2 had more experience solving problems involving averages (item 1); in addition there is some evidence to suggest that boys are more self-efficacious than girls in general (Pintrich & De Groot, 1990) and in mathematics problem solving contexts in particular (Jungle & Dretzke, 1995). As is reported, item 8 functioned differentially across year levels.

External evidence refers to the extent to which the scores obtained from the measure correlate with other previously validated constructs. In this paper a moderate linear association was evident between mathematics self-efficacy and self-efficacy for statistical literacy. This is as expected as students learn statistical concepts in their mathematics classes and indeed data collection for this study occurred during mathematics classes.

Consequential evidence concerns the future impact that any proposed instrument may have on students who complete the instrument. In this instance it is important that items do not differentiate between sub-groups of students (Smith, 2001). As reported, item calibrations for males and females were not significantly different at the 1% level. There is evidence, however, to suggest that item 8 differentiated between younger and older students.

Conclusion

The establishment of validity is an argument that requires a research program rather than a single empirical study (Kane, 2006). In this paper evidence based on the responses of just two samples of students has been presented. This evidence, however, does suggest that the proposed subscale is a valid measure of middle school student's self-efficacy for statistical literacy. It is acknowledged, though, that further item development is warranted, in particular the noted difficulties associated with items 1 and 8.

Acknowledgements

Thanks are extended to Assoc. Prof. Rosemary Callingham and Prof. Jane Watson for their kind assistance and advice. The research was funded by Australian Research Council grant number LP0669106, with support from the Australian Bureau of Statistics.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Bandura, A. (1997). *Self-efficacy: The Exercise of Control*. New York: W.H.Freeman.

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review, 15*(1), 1-40.

Callingham, R., & Watson, J. M. (2005). Measuring Statistical Literacy. *Journal of Applied Measurement, 6*(1), 1-29.

Carmichael, C. S., & Hay, I. (2008). Middle school students' interest in statistical literacy. In M. Goos, R. Brown & K. Maker (Eds.), *Proceedings of the 31st Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 109-115). Brisbane: MERGA.

Gal, I. (2003). Teaching for statistical literacy and services of statistics agencies. *The American Statistician, 57*(2), 80-84.

Holmes, P. (1986). A statistics course for all students aged 11-16. In R. Davidson & J. Swift (Eds.), *Proceedings of the 2nd International Conference on Teaching Statistics* (pp. 194-196). Victoria (BC): IASE.

Jungle, M. E., & Dretzke, B. J. (1995). Mathematical self-efficacy gender differences in gifted/talented adolescents. *Gifted Child Quarterly, 39*(1), 22-26.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education and Praeger Publishers.

Linacre, J. M. (2006). WINSTEPS Rasch measurement computer program (Version 3.61.2) [Computer Software]. Chicago: Winsteps.com.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1-27.

Ma, X., & Kishor, N. (1997). Assessing the relationship between attitude towards mathematics and achievement in mathematics: A meta-analysis. *Journal for Research in Mathematics Education, 28*(1), 26-47.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741-749.

National Curriculum Board. (2008). *National Mathematics Curriculum: Framing paper*. Retrieved November 25, 2008 from: http://www.ncb.org.au/verve/_resources

Panksepp, J. (2003). At the interface of the affective, behavioral, and cognitive neurosciences: Decoding the emotional feelings of the brain. *Brain and Cognition, 52*(1), 4-14.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-40.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*(2), 261--288.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

Trautwein, U., Ludtke, O., Köller, O., Marsh, H. W., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology, 98*(4), 788-806.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals.* Mahwah, NJ: Lawrence Erlbaum Associates.

Watt, H. M. G. (2008). A latent growth curve modeling approach using an accelerated longitudinal design: The ontogeny of boys' and girls' talent perceptions and intrinsic values through adolescence. *Educational Research and Evaluation, 14*(4), 287-304.