Using Rasch Measurement to Identify Cross-cultural Aspects of Statistical Literacy

Rosemary Callingham
University of Tasmania
<Rosemary.Callingham@utas.edu.au>

Statistical literacy requires not only mathematical skills and understanding but also the capacity to apply these and interpret findings in context. Hence instruments developed to measure statistical literacy, which rely on context, might not work effectively in cross-cultural settings. The performances of students in Hong Kong on a test of statistical literacy were compared with archived data from Australian students. Differences in the order of item difficulty between the two groups shed light on the impact of the different cultural settings. Equating and calibrating data from both groups in a single analysis provided indicated that the same construct was being measured across the two cultural settings. Finally, DIF analysis indicated items that functioned differently in the two settings.

The rise of the knowledge economy (Drucker, 1969) places a heavy emphasis on understanding and using statistical information in appropriate ways. Interpreting statistical information requires a grasp of mathematical ideas as well as context and these two aspects can be combined into the construct of statistical literacy.

There are numerous definitions of statistical literacy. Wallman (1993) describes it succinctly as

> 'Statistical Literacy' is the ability to understand and critically evaluate statistical results that permeate our daily lives–coupled with the ability to appreciate the contribution that statistical thinking can make in public and private, professional and personal decisions (p. 1).

Gal (2002) further suggested that the Statistical Literacy required by society was composed of two components:

> (a) people's ability to interpret and critically evaluate statistical information, data-related arguments, or stochastic phenomena, which they may encounter in diverse contexts, and when relevant (b) their ability to discuss or communicate their reactions to such statistical information, such as their understanding of the meaning of the information, their opinions about the implications of this information, or their concerns regarding the acceptability of given conclusions. (pp. 2-3)

These definitions suggest that any assessment of statistical literacy must include aspects of the mathematical underpinnings of statistics, and opportunities to display understanding in context.

Context, however, may be problematic. Social and educational contexts vary from place to place, and inferences about understanding may not translate from one situation to another. This has a potential impact on assessment. Not only may students interpret assessment items in unintended ways, but also the test or assessment instrument itself may be biased. In an increasingly globalised world, however, it is important to be able to interpret assessment information in different places and for diverse purposes. International students, comparative studies and a desire to use world's best practice all combine to demand the highest standards of assessment information. If the assessment process cannot be transferred to a range of educational settings then it is likely to have limited utility.

Many researchers have focussed on tracking students' developing understanding of specific aspects of statistics and probability, or chance and data within the school curriculum (see Watson, 2006 for a summary). Many of these studies have relied on surveys backed up by interviews. Others have developed instruments to measure statistical understanding using testing procedures, often at college or university level, such as those used in the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project (Garfield, del Mas & Chance, 2006). Watson and Callingham (2003) used archived survey data from Australia and Rasch (1960) measurement techniques to develop a scale of statistical literacy at the school level. This scale was replicated in a second Australian study that used some of the original items together with new items with the aim of developing an instrument that could be used by teachers to identify their students' achievement in statistical literacy (Callingham & Watson, 2005).

The Statistical Literacy Scale has six levels that describe the increasing sophistication of students' thinking. The context plays an important part in the descriptor for each level. A summary of the levels is presented in Table 1.

Table 1.
Statistical literacy construct (Watson & Callingham, 2003).

Level	Brief characterisation of levels
6 Critical Mathematical	Critical, questioning engagement with context, using proportional reasoning particularly in media or chance contexts, showing appreciation of the need for uncertainty in making predictions, and interpreting subtle aspects of language.
5 Critical	Critical, questioning engagement in familiar and unfamiliar contexts that do not involve proportional reasoning, but which do involve appropriate use of terminology, qualitative interpretation of chance, and appreciation of variation.
4 Consistent Non-critical	Appropriate but non-critical engagement with context, multiple aspects of terminology usage, appreciation of variation in chance settings only, and statistical skills associated with the mean, simple probabilities, and graph characteristics.
3 Inconsistent	Selective engagement with context, often in supportive formats, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas.
2 Informal	Only colloquial or informal engagement with context often reflecting intuitive non-statistical beliefs, single elements of complex terminology and settings, and basic one-step straightforward table, graph, and chance calculations.
1 Idiosyncratic	Idiosyncratic engagement with context, tautological use of terminology, and basic mathematical skills associated with one-to-one counting and reading cell values in tables.

Despite the growing interest in statistical thinking worldwide, there have been few cross-cultural studies specifically focussing on statistical literacy. Garfield (2003) reports on a study by Lui (1998) that compared similar groups of American and Taiwanese students on a test of statistical reasoning. Several studies of adult literacy, most recently in

2003, have used statistical items in domains such as "document literacy" that could be interpreted as statistical literacy (National Centre for Education Statistics, 2005). There appear to be few cross-cultural studies of statistical literacy reported at the school level, however. Both PISA (Organisation for Economic Cooperation and Development, 2004) and TIMSS (Mullis, Martin, & Foy, 2005) provide cross-cultural measures of mathematical competence in school students, although these are somewhat different. Both these international surveys include some aspects of chance and data from the school curriculum, but neither specifically addresses statistical literacy.

In both of the school-level international studies, students from Hong Kong are among the highest achieving in the world in mathematics. Various reasons have been advanced for this success including the Confucian tradition (Leung, 2005). Chinese students aim to achieve perfection through close imitation and hard work (Li, 2004). Such approaches, however, may not lead to success on statistical literacy tasks where inference and application in context may demand interpretation as well as mathematical skills.

If issues of cultural differences in statistical literacy are to be explored, it is necessary to establish dependable measures of statistical literacy in a variety of situations. The Statistical Literacy Scale (Watson & Callingham, 2003) had been validated in different ways in Australian settings, but had not been tested in different countries, especially Asian settings. With this in mind, the research questions for the study are

1. Does the Statistical Literacy Scale provide valid information about students' performances in the different cultural setting of Hong Kong?
2. If the scale is working appropriately, what other factors affect students' performances in Hong Kong?

This report presents the initial findings from a larger study undertaken in Hong Kong across middle years' grades. Results from a second year high school grade were compared with an archived data set from Australia. The Hong Kong school was an English medium private school that followed the English curriculum. The inclusion of this school allowed a consideration of the ways in which the statistical literacy items performed in a different social context but without the added complexity caused by translation, because the instrument could be administered in English.

Methodology

A test of Statistical Literacy comprising 28 items that had been validated in previous studies (all items are detailed in Callingham & Watson, 2005 and Watson & Callingham, 2003) was administered to 195 students in all second year high school classes in one English speaking high school in Hong Kong. The mean age of the students was 11.86 years with a range of 11 to 13 years, and the students were in their first term of the new school year. Of the cohort, 112 (57.7%) students were male, 78 (40.2%) female and four (2.1%) did not record their gender. The relatively high proportion of males is consistent with the social context of Hong Kong where male children are favoured by parents, and hence are more likely to attend prestigious schools. Although the social grouping was not formally recorded, discussion with the school principal indicated that a majority of students came from Hong Kong Chinese backgrounds, often where the parents and children also held other citizenship, a shift that had occurred since Hong Kong became a Special

Administrative Region of China. Prior to this event, the school population had been mainly Western expatriate students. This latter group was now in a minority in the school, and a number of other groups, particularly from Korean and Indian backgrounds, were also represented.

All school subjects were taught in English, usually by expatriate teachers. The students were asked whether they spoke English at home, and how long they had lived in Hong Kong. The summaries are shown in Table 1 and Table 2.

Table 1.
Language spoken at home by sex

	English		Chinese		Other	
M	82	(73.21%)	22	(19.64%)	8	(7.14%)
F	47	(60.26%)	14	(17.95%)	17	(21.79%)
Unknown	2	(50.00%)			2	(50.00%)
Overall	131	(67.53%)	36	(18.56%)	27	(13.92%)

Table 2.
Length of time students lived in Hong Kong by sex

	< 1 year		Between 1 and 5 yrs		More than 5 years		All my life	
M	5	(4.42%)	19	(16.81%)	37	(32.74%)	52	(46.02%)
F	8	(10.26%)	11	(14.10%)	25	(32.05%)	34	(43.59%)
Unknown	0		1	(25.00%)	1	(25.00%)	2	(50.00%)
Overall	13	(6.67%)	31	(15.90%)	63	(32.31%)	88	(45.13%)

In summary, the Hong Kong students had lived in the country for most of their lives, mainly spoke English at home, and appeared to come from a relatively privileged social grouping from Chinese or other Asian backgrounds.

For comparison purposes, a data set was created from archived data. The 210 students were in Grade 7 or 8 from one Catholic high school in Tasmania, and had undertaken a test of Statistical Literacy as part of a middle years study (Callingham & Watson, 2005). The breakdown of the group of Tasmanian students by sex and grade is shown in Table 3.

Table 3.
Characteristics of the Australian group created for comparison purposes

	Sex				Total	
	M		F			
Year 7	59	(52.68%)	53	(47.32%)	112	(53.33%)
Year 8	59	(60.20%)	39	(39.80%)	98	(46.67%)
Total	118	(56.19%)	92	(43.81%)	210	

The Tasmanian group was more balanced across gender than the Hong Kong group. Although no age range was available from the Australian data, students across Year 7 and Year 8 in Tasmania could be expected to have a similar age range to students in Hong Kong. All Australian students spoke English as a first language. No data were collected

about the length of time the Australian students had lived in the country, but Tasmanian demographics suggest that a majority of them had lived in Australia, and Tasmania, for most of their lives.

The purpose of the study was principally to consider the behaviour of a scale of statistical literacy in the different cultural context of Hong Kong, rather than to compare the performances of students in Hong Kong and Australia. The items that were used in Hong Kong had been used in several Australian studies (Watson, Kelly & Izard, 2006), and the interest was in identifying whether the cultural context in which the items were presented affected their behaviour. The groups were not random or representative samples in either country, but had been chosen for convenience. The students included in each group, however, were immersed in very different social settings, and came from diverse cultural backgrounds. As such they provided two diverse populations of students in order to test the instrument used.

Data Analysis

Responses to the Hong Kong test were coded using scoring rubrics that had been validated in prior Australian studies. The rubrics were based on structural complexity and statistical appropriateness of the responses (Callingham & Watson, 2005). The Australian data had been previously coded and the archived data from the target grades were used to create an Australian data set for comparison purposes. The coded data were then analysed using the Rasch Partial Credit Model (Masters, 1982) using Quest computer software (Adams & Khoo, 1996). Hong Kong students completed 28 items. Australian students attempted 49 items across two different linked test forms. Of these item sets, 15 items were common across both groups of students. Several different analyses were undertaken to reveal diverse item behaviour across the two groups.

1. Separate analyses were undertaken for Hong Kong (HK) and Australian (OZ) data, unlinked and unanchored. The aim of this analysis was to consider changes in the order of difficulty of the items. The fit to the model was also examined to ensure that the items were working together consistently within each sample. Although the Rasch model is generally considered to be sample free it is not population free (Bond & Fox, 2001). Large shifts in the order of item difficulty could indicate that the instrument was not appropriate to use across populations, even when it provided consistent information within a population. Reliability measures were also obtained for both items and persons.

2. A combined HKOZ data set was created and the two item sets were equated and calibrated in a single operation (Griffin & Callingham, 2006). Item fit and difficulty were again examined to provide evidence that the scale behaved consistently across two different cultural groups. If this was the case, then the scale could be used in future studies to provide comparisons of performance between diverse cultural groups.

3. Finally a differential item functioning (DIF) analysis was completed using the combined data set using PLACE as the grouping variable. This analysis provided information about whether or not the items were working in the same way across

both groups. Significant differences between groups would indicate item bias that might occur because of differences in curriculum or context (Bond & Fox, 2001). At the test level, overall measures of DIF provided information about bias in the test. Even if there was some item level DIF, if this was balanced across the groups then the overall test would not show bias and could be used to make valid inferences about each group, and provide comparison measures.

Results

Separate unanchored analyses

In both Hong Kong and Australian analyses there was no serious item misfit. One item (TRV3) showed some randomness or underfit in the Hong Kong data (Infit Mean Square = 1.48, Infit t = 4.47; Outfit Mean Square = 1.64, Outfit t = 4.43). This item required an inference to be drawn and justified on the basis of data about travel to school presented in a pictogram. Two items (SPT4, RASH) showed randomness in the Australian data. SPT4 (Infit Mean Square = 1.39, Infit t = 1.69; Outfit Mean Square = 2.09, Outfit t = 2.76) was not used in Hong Kong and also involved drawing an inference and justifying this on the basis of data about a sporting contest presented in a table. RASH (Infit Mean Square = 1.53, Infit t = 3.91; Outfit Mean Square = 1.99, Outfit t = 3.49) was used in Hong Kong and involved students making a decision about the meaning of the risk of developing a rash on the basis of evidence on a medicine bottle. No other misfit was observed in either data set, indicating that the items generally appeared to be working consistently within both samples to provide a measure of statistical literacy.

When the order of difficulty of each set of items was compared, again differences were minor. The easiest and hardest items for each group of students were compared. Of the five easiest items for each group only two items did not appear in both sets. Similarly, of the six most difficult items, two were not common to both groups. Details are shown in Table 4. The items appearing only in one list are shown in bold font for clarity.

Table 4.
Easy and hard items by PLACE .

	Easy Items			Hard Items	
	OZ	HK		OZ	HK
Decreasing difficulty ↓	HGT2	HGT2	Increasing difficulty ↑	HSE3	HSE3
	MV10	TRV2.1		HSE1.2	T2X2.5
	HGT1	**RASH**		T2X2.5	**RAND.3**
	RAND.1	HGT1		**AMEA.4**	TRV3.5
	TRV2.1	RAND.1		TRV2.3	TRV2.3
				TRV3.5	HSE1.2

Item and person separation reliability measures produced by Quest software were also considered. These provide evidence of the spread of the items or persons along the underlying variable, and have an ideal value of 1. For both items (HK, 0.88; OZ, 0.89) and cases (HK, 0.81; OZ, 0.79) the values were high suggesting that the instruments used provided reliable measures of statistical literacy in both situations.

These findings appear to indicate that for each group of students the items produced a dependable scale of statistical literacy that could be used as a basis for drawing inferences about performance within each place. There was little indication of severe curriculum or contextual influence, since the order of item difficulty appeared to be similar within each group.

Combined data sets analysis

The overall fit to the model of the data from the combined data set was acceptable with the Infit Mean Square at the ideal value of 1.00, and only one item (SPT4 given only to Australian students) showed any misfit. The items were working effectively across the two combined groups to provide a single measurement scale, Statistical Literacy. Item separation reliability (0.95) and person separation reliability (0.86) were high indicating that a satisfactory metric was obtained. Combining these two different groups of students into a new population did not appear to affect the behaviour of the items. This would suggest that the scale could provide dependable measures for comparisons across countries.

DIF analysis

The final analysis compared the ways in which the items behaved across the two groups from Hong Kong and Australia. The results are presented graphically as a plot of standardised differences between the groups, rather than logit values. The positive or negative sign is associated with one of the two comparison groups and significant differences are those having values < -2 or >2. Hence a visual picture of items that the different groups found easier is provided. Of the common items, 13 provided meaningful DIF measures and these are shown in Figure 1. Items significantly easier for Australian students were RAND, T2x2 and TRV3. Those significantly easier for Hong Kong students were AMEA and AOUT.

	Easier for OZ								Easier for HK						
	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
rand			*			.		|		.				
hse1						.		|	*	.				
hse2						.		|	*	.				
hse3						.		|	*	.				
t2x2					*	.		|		.				
hgt1						.	*	|		.					
hgt2						*.		|		.				
mv10						.	*	|		.					
mv11						.		|		*.				
trv3		*				.		|		.				
var						.	*	|		.					
amea						.		|		.				*	
aout						.		|		.				*	

Figure 1. DIF measures by PLACE.

Discussion

The findings from this initial study of statistical literacy in Hong Kong and Australia suggest a number of points of interest. The social context of Hong Kong did not appear to play a part in the ways in which students approached items contextually based in Australia. For example, the three items HSE1, HSE2 and HSE3 were based on a media report of median house prices in Hobart, Tasmania. Students were required to explain the terms average and median, used in the report, and explain why the median was used. It is noticeable that these items showed no significant DIF, and were in fact slightly easier for Hong Kong students. HSE3, explaining why the median was used, was the most difficult item for both groups, suggesting that use of the median is not well covered in either curriculum. It seems that relevance does not necessarily require a basis in the identical social context in which the students live, but more addresses social situations with which they can identify. House prices are as much an issue in Hong Kong as they are in Australia, and adolescents are aware of their importance.

The items used provided a dependable and unbiased scale for both groups of students, indicating that they can be used with confidence in different settings. The Asian setting of Hong Kong is very different from that of Tasmania, but the scale was stable, showed almost no misfit and good reliability in both contexts. When the groups were combined, the scale was also well behaved, providing reliable estimates for item difficulty and person ability. The scale of statistical literacy appears to be the same construct for both Hong Kong and Australian students, indicating that genuine comparisons could be made. These findings are important for international comparisons. If a dependable scale can be obtained then the way is open for further studies that consider other factors that may impact on achievement in statistical understanding, such as teaching approaches and curriculum emphasis.

There are hints of diverse curriculum emphases in the DIF results. The RAND item, found easier by Australian students, asked students to explain what “random” meant. It was coded with an emphasis on statistical appropriateness and RAND was among the most difficult items for Hong Kong students. The Hong Kong students’ work indicated a lack of familiarity with this kind of item and they were likely to suggest colloquial meanings for the word rather than to draw on statistical ideas. It is possible, also, that language played a part, despite the school being an English-medium school. TRV3, also found easier by Australian students, presented students with data in a pictogram about how students travel to school. The modes of transport included bus, train, car, walk and cycle. Students were required to suggest how an absent student would get to school the next day, and to justify their answers. Highest levels of response were able to include the implicit uncertainty but make sensible suggestions based on the data. This inference drawing exercise appeared more difficult overall for the Hong Kong students, although the highest level of response was in the top six list of difficulty for both groups. The final item found easier by Australian students was T2x2, in which data about the incidence of lung cancer among smokers and non-smokers was presented. This was also difficult for both groups at the highest level, which required responses that considered all the data and applied proportional reasoning to identify that the incidence was the same. It seems unlikely that Australian students are better able to reason proportionally than students in Hong Kong, but they may be more likely to experience data presented in two way tables.

In contrast, the items found easier by the Hong Kong students were the two addressing average, AMEA and AOUT. In the context of a science experiment, in which a selection of results including an obvious outlier were given, students were asked to respond to different suggestions for calculating the mean (AMEA) or dealing with the outlier (AOUT). AMEA was in the most difficult list for Australian students, and it was noticeable when marking the Hong Kong students' work that they appeared to have a sophisticated understanding of the mean and could critique ways of dealing with the outlier. This may reflect a curriculum emphasis on central tendency, but also some knowledge transfer from science classes.

No attempt was made in this initial analysis to place the Hong Kong students into levels on the Statistical Literacy Scale. It appears, however, that the instrument used is reliable and valid, and that the scale interpretation would be the same across these different groups of students.

Implications

The use of the Statistical Literacy Scale in an Asian context is an important development. Little work has been done on statistical literacy in cross cultural settings. Although this study reports only on English speaking students, the student profile suggests that most of the Asian students had spent nearly all their lives in Hong Kong, growing up in an Asian cultural setting. Data have been collected from Cantonese speaking students and analyses of these data will provide additional insights. The Statistical Literacy Scale provided a valid and reliable measure in both Australia and Hong Kong. This indicates that, with a robust instrument available, future work could focus on cultural and social context differences in the area of statistics.

The suggestion from the high mean score of students who spoke Cantonese at home that Chinese students are performing in ways similar to those seen in international studies of mathematical performance deserves further investigation. The capacity to undertake cross-cultural studies using sound instruments of aspects of learning different from those addressed in large scale international studies provides potentially new avenues of research, in this case focussing on a relatively new area of the curriculum. This initial study has confirmed the applicability of the Statistical Literacy Scale in new cultural contexts and opens the way for further intriguing research.

Finally, it is worth noting the utility of Rasch measurement in this study. The use of Rasch measurement approaches provided several ways in which the data could be examined. Not only did it provide quality control of the instrument used, but it also allowed consideration of differences in item difficulty across two populations, Hong Kong and Australia. In addition, DIF analysis was used to explore bias in the items and to provide insight into potential educational differences. Through the use of equating, Rasch measurement also afforded measures of ability that could be used to make direct comparisons using conventional statistical techniques. The capacity to consider different aspects using the single technique is a testament to the power and flexibility of the Rasch measurement approach.

Acknowledgement

This study was supported by a Cheung Kong Endeavour Fellowship undertaken at the Hong Kong Institute of Education.

References

Adams, R.J. & Khoo, S. (1996). *Quest: The interactive test analysis system, Version 2.1*. [Computer software]. Melbourne: ACER.

Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch model. Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum.

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

Drucker, P. (1969). *The age of discontinuity: Guidelines to our changing society*. New York: Harper and Row.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review,* 70, 1-51.

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22–38.

Garfield, J., del Mas, R. & Chance, B. (2006). *About the ARTIST project*. Accessed 21 March 2007 from https://app.gen.umn.edu/artist/about.html

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

Leung, K. S. F. (2005, August). *In the books there are golden houses: Mathematics assessment in East Asia.* Plenary address to the ICMI 3rd East Asian Regional Conference on Mathematics Education, Shanghai.

Li, J. (2004). A Chinese cultural model of learning. In L. Fan, N-Y, Wong, J. Cai, & S. Li (Eds.) *How Chinese learn mathematics: Perspectives from insiders* (pp. 124-156). Singapore: World Scientific Publishing.

Lui, H. J. (1998). *A cross-cultural study of sex differences in statistical reasoning in college students in Taiwan and the United States*. Unpublished doctoral dissertation. Minneapolis: University of Minnesota.

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

Mullis, I.V.S., Martin, M.O., & Foy, P. (2005). *IEA's TIMSSS 2003. International report on achievement in the mathematics cognitive domains. Findings from a developmental project.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Accessed 25 March 2007 from http://timss.bc.edu/timss2003i/mcgdm.html

National Centre for Education Statistics (2005). *Highlights from the 2003 International Adult Literacy and Lifeskills survey (ALL).* Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Organisation for Economic Cooperation and Development (OECD) (2004). *Learning for tomorrow's world. First results from PISA 2003*. Paris: Author.

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).

Wallman, K. K. (1993) Enhancing Statistical Literacy: Enriching our society. *Journal of the American Statistical Association, 88*(421), 1–8.

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

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

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