The Vocabulary of Statistical Literacy

Jane M. Watson and Ben A. Kelly

University of Tasmania

Abstract

This paper considers the development of school students' understanding of three terms that are fundamental to statistical literacy: sample, random, and variation. A total of 738 students in grades 3, 5, 7, and 9 were asked in a survey to define and give an example for the word sample. Of these, 379 students in grades 7 and 9 were also asked about the words random and variation. Responses were used to describe developmental levels overall and to document differences across grades on the understanding of these terms. Changes in performance were also monitored after lessons on chance and data emphasising variation for 335 students. After two-years, 199 of these students and a further 209 students who were surveyed originally but did not take part in specialised lessons, were surveyed again. The difference after two-years between the performance of students who experienced the specialised lessons and those who did not was considered, revealing no differences in longitudinal performance. For students in grades 7 and 9 the association of performance on the three terms was explored. Implications for mathematics and literacy educators are discussed.

Introduction

Although literacy and numeracy are often linked in governmental and education department policy statements on objectives for students in Australian schools, numeracy comes out as the poor relation when it comes to funding priorities and the attention of teachers across curriculum areas in schools (Department of Education Tasmania, 2000; Bridge, 2002). In preservice teacher education programs, moves to courses on multiliteracies have not enhanced the prospects for numeracy greatly, in that so many different literacies are demanding equal billing with "literacy" literacy. Searches of the literature on multiliteracies produce no mention of numeracy, while focusing much attention on other areas such as information technology literacy, cultural literacy, civic literacy, and drama literacy. Even efforts in some countries to talk about quantitative literacy rather than numeracy have not done very much to enhance the stance of the quantitative side of literacy. For those interested in statistical literacy as a subset of quantitative literacy, the hope to attract the attention of the multiliteracies advocates appears even more remote.

Advocates of statistical literacy point to definitions such as the one presented by the 1992 president of the American Statistical Association.

> 'Statistical Literacy' is the ability to understand and critically evaluate statistical results that permeate our daily lives – coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions. (Wallman, 1993, p.1)

Such a definition requires a great deal of statistical understanding and some might claim a course in statistics. Is statistical literacy really a legitimate multiliteracy that should be expected of school students, and their teachers?

In answering "yes" to this question, the authors note that the emphasis in statistical literacy on critical thinking and evaluation and on contributions to public and private life are similar to the goals of literacy itself, stated for example by Luke and Freebody (1997) with respect to reading. When statisticians view statistical literacy, however, they have a tendency to focus more on the "statistical" than the "literacy." This can be seen in the procedural aspects of finding probabilities of events, drawing graphs, and calculating mean values. The literacy aspects of communicating outcomes in words and being able to describe the concepts involved in ideas like "chance" or "average," often receive little or no attention. This is seen for example when students are asked to define or describe what average means and all they can say is, "it's when you add-um-up and divide." It is also seen in cold conclusions like "$t = 4.2$, $p < .01$, so the result is statistically significant."

This study considers the dilemma of statistical literacy and its place at the intersection of statistics and literacy. Statistics requires the basic understanding of statistical concepts, such as random, whereas literacy requires the ability to express that understanding in words not in mathematical formulas. If statistical literacy is to have any prospect of attracting attention from the multiliteracies community, then it must demonstrate an expectation for literacy skills within the field. To do this there must be an expectation that students can describe the meaning of the concepts that are the foundation of statistical thinking. With the exception of the work done by Malone and Miller (1993) and Miller (1993) on mathematical vocabulary, there has been little research into school students' expression of their understanding of mathematical or statistical concepts.

Until very recently the measurement of statistical understanding has focused on procedures with numerical answers and not on descriptions of terminology. Even Malone and Miller (1993; Miller 1993), who focused on mathematical vocabulary, gave partial credit to students who responded with symbols or a combination of symbols and explanations. The current study grew out of earlier research that included "definitions" as part of a survey of student understanding of chance and data (Watson, 1994). Watson, Collis, and Moritz (1993) reported on preliminary data for the following questions, asked to students in grades 3, 6, and 9.

- If someone said you were "average", what would it mean?
- What things happen in a random way?
- If you were given a "sample", what would you have?

Moritz, Watson, and Pereira-Mendoza (1996) later reported on the comparison of responses from Australian and Singaporean students on these questions. For the Australian data set of 1014 students, 40% of grade 3 students did not respond to the question on average, but this reduced to 7% of grade 6 and 2% of grade 9. For the random question, non-response occurred for 78% of grade 3 students, 40% of grade 6, and 16% of grade 9. For the sample question the corresponding non-response rates were 42%, 16%, and 7%. On the other hand, the idea of average as the centre of a data set (not just "centre") was suggested by 6% of grade 3, 38% of grade 6, and 62% of grade 9. The description of two or more aspects of the process of things happening in a random way, beyond giving examples, was accomplished by 1% of grade 3, 3% of grade 6, and 19% of grade 9. For sample, descriptions equivalent to "a small part representing a whole" were given by 3% of grade 3, 15% of grade 6, and 25% of grade 9. Overall students were most familiar with the term average and least familiar with random.

Watson and Moritz (2000) analysed the question on sampling further in a longitudinal study of a sampling subscale of items. For 151 grade 3 students who answered the question again after two years and after four years, the rate of non-engagement with the idea of sample dropped from 50% to 26% (in grade 5) to 12% (in grade 7), whereas the rate of response describing two features of a sample rose from 11% to 33% to 49%. The rate of response including all basic aspects of a sample rose from 3% to 7% to 16%.

The current study used variations on two of the above items, on random and sample, and replaced the item on average with one on variation, as reported in Watson, Kelly, Callingham, and Shaughnessy (2003). The questions are shown in Figure 1. In each case a direct approach of asking for the meaning of the term is used. Examples then are requested and for variation, students are asked to put the word in a sentence. This approach was used to give the greatest opportunity for students to display understanding and to reduce the non-response rate.

1a) What does "sample" mean?
1b) Give an example of a "sample".

2a) What does "random" mean?
2b) Give an example of something that happens in a "random" way.

3a) What does "variation" mean?
3b) Use the word "variation" in a sentence.
3c) Give an example of something that "varies".

Figure 1. Questions used in Watson et al. (2003)

The analysis of Watson et al. (2003) produced a four-step coding scheme for each item. This scheme is based on two aspects of the overall responses to the parts of each question: the structural complexity and the statistical appropriateness. The complexity was judged in relation to a developmental model (Biggs & Collis, 1982, 1991) where elements related to the task were noted. A code of 0 was given for tautological responses or those that did not show any appreciation of meaning. A code of 1 was given for responses reflecting a single idea related to the term or an example. Code 2 responses reflected a straightforward explanation of the term and an example. Code 3 responses gave more complete explanations, related to the examples given. Illustrations of responses for the codes for each term are given in Table 1.

Table 1
Examples of Responses from Watson et al. (2003)

Coding	Sample	Random	Variation
0	'The number of something' (28.9%)	'Very quickly' (40.3%)	'Jayden varies from place to place' (41.7%)
1	'Try something. Getting a sample of water.' (34.4%)	'Choosing something. Random breath test.' (19.9%)	'You get a choice. A car varies from sizes and colours.' (25.6%)
2	'You have a small piece of something. I had a sample to	'It means in any order. The songs on the CD came out	'How something changes. The weather.' (25.1%)

	eat at the supermarket.' (23.6%)	randomly.' (36.5%)	
3	'Take a small amount of one thing away to test. Blood sample.' (13.1%)	'Random means something that does not happen in a pattern. In a Tatts lotto draw …' (3.3%)	'Slight change or difference. People vary in size …' (7.6%)

Table 1 provides the overall percent of responses in each code for each question for the Watson el al. (2003) study. Watson et al. also performed a partial credit Rasch analysis on the overall scale, which measured students' understanding of chance and data with an emphasis on variation. This analysis produced a variable map that estimated student abilities and item difficulties on the same scale. The underlying construct, representing understanding of variation within a chance and data context, was then considered in a developmental sense based on the expectations of the items and their placement along the scale. Four levels of increasing understanding were identified. At Level 1, Prerequisites for Variation, items upon which students were likely to be successful were associated with working out the environment, reading simple graphs and tables, and reasoning intuitively about chance. At Level 2, Partial Recognition of Variation, items were associated with putting ideas in context, focussing on some single aspects of tasks while neglecting others. At Level 3, Applications of Variation, items were associated with consolidating and using ideas in context but being inconsistent in picking the most salient features. At Level 4, Critical Aspects of Variation, items required employing of complex justification or critical reasoning. Although this analysis was based on a total of 44 individual partial credit items, some as parts of larger tasks, the items used in the current study, contributed 9 items, 3 based on the codes 1 to 3, for each of the three questions in Figure 1 (see Table 1). In the case of these three questions, the analysis and determination of thresholds for the levels placed the Code 1 category of response for each question at Level 2, the Code 2 category at Level 3 for each question, and the Code 3 category at Level 4 for each question.

As the analysis of Watson et al. (2003) was based on the complete data set, regardless of grade, it was of interest to consider performance for the different grade cohorts on the three questions. Other research questions arose from the larger project within which this study was placed.

Research Questions

The research questions arise in the context of concern about the linguistic aspects of literacy within statistical literacy and the opportunity to observe change and association within a larger study.

1. What levels of understanding are displayed for the term "sample" in grades 3, 5, 7, and 9, and for the terms "random" and "variation" in grades 7 and 9?

2. What changes in levels of understanding occur for students who have experienced a unit of instruction on chance and data emphasising variation?

3. What changes occur in levels of understanding of the terms over a period of two years? Is there any difference for students who have experienced the unit of instruction after the first survey?

4. For students in grades 7 and 9 what is the degree of association between levels of response for the pairs of terms?

Methodology

Sample

The sample of students who participated in this study will be described as Sample 1, Sample 2A, Sample 2B, and Sample 3. The students in Sample 1 were in grades 3, 5, 7, and 9 from 10 government schools in the Australian state of Tasmania. A subset of the students from Sample 1 who had completed specialised lessons focusing on variation in the chance and data curriculum and who had also completed a post-test survey including the questions in Figure 1 made up Sample 2A. A further subset, Sample 2B was again re-tested two years later with the same survey. Sample 3 was a disjoint subset of students from Sample 1, matched by school for socio-economic status. Students from Sample 3 were re-tested two years later with the same survey but did not complete the specialised lessons. Students in Samples 2B and Sample 3 were in grades 5, 7, 9, and 11 when they were re-tested after two years. The numbers of students in each sample and grade are shown in Table 2.

Table 2
Number of Students per Grade and Sample

	Grade 3	Grade 5	Grade 7	Grade 9	Grade 11	Total
Sample 1	176	183	186	193		738
Sample 2A	72	82	91	90		335
Sample 2B		47	57	67	28	199
Sample 3		67	44	67	31	209

Procedure

The research team and classroom teacher administered the questions in Figure 1 as part of a larger survey (Watson et al., 2003). Surveys were competed by the students in class time and the entire survey took approximately 45 minutes. For the students in Sample 2A who experienced specialised lessons on chance and data, students in grades 3 and 5 were taught ten lessons of work focusing on variation by an experienced teacher supplied by the research team. Details of these lessons are outlined in Watson and Kelly (2002a). Students in grades 7 and 9 were taught a unit of work that focused on variation by their usual mathematics teacher. The unit of work was devised and supplied by the research team. Details are given in Watson and Kelly (2002b). For the longitudinal follow-up that occurred two years later, there was some difficulty in locating students who were originally in grade 9 and had then changed schools for grade 11. As a result, only 26% of grade 9 students were re-tested.

Analysis

The questions in Figure 1 were analysed using the four-step coding scheme described earlier and used in Watson et al. (2003). The responses from Sample 1 will be used to discuss the distribution of levels of responses across grades for the three questions in Figure 1. *T*-tests were performed to compare the cohorts in different grades in Sample 1. These values are considered to be indicative of change since random assignment of students could not occur. Paired *t*-tests were used to indicate whether there was an increase in performance for Sample 2A after the series of specialised lessons on chance and data and for Sample 2B after a further two years. Paired *t*-tests were also used to gauge the rate of improvement for students in Sample 3 who did not receive any specialised lessons. Difference scores were compared for individual grades across Samples 2B and 3 with *t*-tests to determine if there were differences depending on exposure to instruction.

The large number of tests carried out in the study (approximately 50) suggests that the Bonferroni correction be applied and that statistical significance be considered at the .001-level. In light of this and the information provided by the large number of *p* values less than .05, all *p*-values will be reported for consideration.

Results

Research Question 1: Understanding across Grades

Table 3 shows the percent of students responding at each code for each grade for each of the questions. As can be seen there was an increase in performance with grade for "Sample." More students in grades 7 and 9 responded at Code 3 than did students in grades 3 and 5. With the exception of students in grade 3, there was not very much change in performance across grades at Code 1 but there was an improvement from grade 7 to 9 at Code 2. For "Random" and "Variation" there was little change in performance between grade 7 and 9. Grade 9 students performed marginally better on "Variation" than grade 7 students.

Table 3
Percents of Students per Grade for Each Code in Sample 1

Code		0	1	2	3
Sample	Grade 3	67.1	24.4	6.8	1.7
	Grade 5	30.1	48.6	18.0	3.3
	Grade 7	36.6	36.0	17.2	10.2
	Grade 9	29.5	33.2	27.5	9.8
Random	Grade 7	52.7	19.3	26.9	1.1
	Grade 9	48.2	20.7	30.1	1.0
Variation	Grade 7	59.7	15.6	21.0	3.7
	Grade 9	50.2	20.2	24.9	4.7

Table 4 shows the mean code and standard error for each grade for each of the questions. *T*-tests for "Sample" showed a relatively large difference between grades 3 and 5, less difference between grades 7 and 9, and no significant difference between grades 5 and 7. There was no significant difference between grades 7 and 9 on "Random" or "Variation." Results of *t*-tests between grades are also shown in Table 4. The actual difference in mean coding level is not great for grades 7 and 9, and only half a code level for grades 3 and 5. This represents approximately 2/3 of a standard deviation improvement across these two grades.

Table 4
Means, Standard Errors, and t-tests for Sample 1

		Grade 3		Grade 5		Grade 7		Grade 9
Sample	Mean	0.43		0.95		1.01		1.18
	Std. Error	0.053		0.058		0.072		0.070
	t, p		-6.55, $p<.001$		-0.71, NS		-1.66, $p<.05$	
Random	Mean					0.76		0.84
	Std. Error					0.065		0.064
	t, p						-0.83, NS	
Variation	Mean					0.69		0.84

Std. Error		0.068	0.069
t, p		-1.56, *p*<.06	

Research Question 2: Change after Instruction

Paired *t*-tests were conducted for students in Sample 2A to measure change in performance after instruction. Means, standard errors, and test results are shown in Table 5. For "Sample" improvement occurred for grades 3, 5, and 7 from the pre- to the post-test, with the greatest improvement for grade 5. For the terms "Random" and "Variation" grade 7 improved on both, with a greater improvement on "Variation." Grade 9 students improved on "Random" only. As can be seen, however, the outcomes represented at most about a third to a half of a coding level improvement for all means.

Table 5
Pre and Post Means, Standard Errors, and Paired t-tests for Sample 2A

		Grade 3	Grade 5	Grade 7	Grade 9
Sample	Pre Mean, Std. Error	0.53, 0.083	0.93, 0.073	1.04, 0.107	1.09, 0.101
	Post Mean, Std. Error	0.83, 0.116	1.44, 0.096	1.32, 0.086	1.29, 0.121
	t, p	-2.42, *p*<.01	-4.49, *p*<.001	-2.25, *p*<.02	-1.42, *p*<.08
Random	Pre Mean, Std. Error			0.79, 0.085	0.81, 0.083
	Post Mean, Std. Error			1.03, 0.087	1.10, 0.095
	t, p			-2.23, *p*<.02	-2.30, *p*<.02
Variation	Pre Mean, Std. Error			0.60, 0.079	0.81, 0.097
	Post Mean, Std. Error			0.97, 0.092	1.03, 0.098
	t, p			-3.48, *p*<.001	-1.77, *p*<.08

Research Question 3: Change after Two Years

The students in Sample 2B, constituted a subset of Sample 2A who had experienced instruction and that completed the longitudinal survey two years later. Table 6 shows the means and standard errors for each of the tests, for each grade and question. The pre-test and post-test data are reported here again because Sample 2B is a subset of Sample 2A and therefore values are not identical. Again, paired *t*-tests were conducted to detect any change in performance across the three tests.

Table 6
Pre, Post and Longitudinal Means, Standard Errors and Paired t-tests for Sample 2B

		Grade 3/5*	Grade 5/7*	Grade 7/9*	Grade 9/11*
Sample	Pre Mean, Std. Error	0.57, 0.125	0.93, 0.089	0.96, 0.120	1.21, 0.215
	Post Mean, Std. Error	0.83, 0.154	1.33, 0.101	1.22, 0.100	1.32, 0.253
	Longitudinal Mean, Std. Error	0.94, 0.104	1.26, 0.107	1.51, 0.124	1.57, 0.202
	Pre-Post Change *t*	-2.01, *p*<.03	-2.95, *p*<.01	-1.71, *p*<.05	-0.41, *p*=.34
	Pre-Longitudinal Change *t*	-2.41, *p*<.02	-2.46, *p*<.01	-4.14, *p*<.001	-1.58, *p*<.07
	Post-Longitudinal Change *t*	-0.70, *p*=.24	0.50, *p*=.31	-1.95, *p*<.03	-1.10, *p*=.14
Random	Pre Mean, Std. Error			0.73, 0.098	0.96, 0.147
	Post Mean, Std. Error			0.94, 0.100	1.14, 0.206
	Longitudinal Mean, Std. Error			1.09, 0.095	1.32, 0.225
	Pre-Post Change *t*			-1.54, *p*<.07	-0.87, *p*=.20
	Pre-Longitudinal Change *t*			-2.95, *p*<.01	-1.63, *p*<.06
	Post-Longitudinal Change *t*			-1.26, *p*=.11	-1.04, *p*=.15
Variation	Pre Mean, Std. Error			0.58, 0.093	1.04, 0.231
	Post Mean, Std. Error			0.90, 0.105	0.93, 0.223
	Longitudinal Mean, Std. Error			1.24, 0.141	1.32, 0.211
	Pre-Post Change *t*			-2.45, *p*<.01	0.42, *p*=.34
	Pre-Longitudinal Change *t*			-5.11, *p*<.001	-1.39, *p*<.09

Post-Longitudinal Change t	-2.60, p<.01	-2.09, p<.03

*Grade in Longitudinal Follow-up

The results in Table 6 corroborate the changes in performance from the pre- to the post-test for grades 3, 5, and 7 on "Sample" found in Sample 2A (Table 5) as well as show a related change in performance from the pre-test to the longitudinal follow-up for these three grades on "Sample." The greatest level of improvement was from the pre-test to the longitudinal follow-up for students who were originally in grade 7, again about half of a coding level. A change between the post-test and the longitudinal follow-up was also evident for grade 7. For "Random" there was a change from the pre-test to the longitudinal follow-up for grade 7 students. There were no significant changes in performance for students originally in grade 9 over any of the tests on "Random." Students originally in grade 7 improved over all three of the testing conditions on "Variation", with a significant change overall from the pre-test to the longitudinal follow-up. Students originally in grade 9 improved on "Variation" between the post-test and the longitudinal follow-up after a non-significant decline in performance from the pre-test to the post-test.

Change in performance is illustrated with examples for each term. For "Sample" a student in grade 3 responded on the pre-test and the post-test without a clear understanding or improvement on the term (Code 0). In the longitudinal follow-up, however, the student then in grade 5 responded with a clearer understanding of the term, recognising the part-whole relationship (Code 2).

- "When you look through something" [pre-test].
- "You could sample things, e.g., You could sample through your hair" [post-test].
- "Choose a couple of things to represent them all, e.g., She chose a sample of books to read" [longitudinal].

Another student, in grade 5 responded to "Sample" on the pre-test (Code 1) and post-test (Code 3) as follows.

- "I think it means sample of perfume or blood" [pre-test].
- "To take a bit of something and try, e.g., make-up at a clinic" [post-test].

This improvement reflects change from only providing an example, to noting the part-whole relationship with the representative aspect of trialing along with a sample. On the longitudinal follow-up, the student sustained the high-level performance from the post-test (Code 3).

- "It means a little bit of something to try, e.g., A tray of new cheese for sampling (you can try)" [longitudinal].

For "Random" the following student in grade 7 improved from the pre-test to the longitudinal follow-up, but not from the pre-test to the post-test, responding at Code 1 on the first two occasions and at Code 2 on the third.

- "To random something means to pick out somebody at random. The man picked his name out of a hat at 'random'" [pre-test].
- "A random pick of something from something, etc. In a Playstation game you can choose a level randomly" [post-test].

- “To pick or choose something without looking at what it is. A cat eats a mouse at random” [longitudinal].

The following student again improved from the pre-test to the longitudinal follow-up, but not from the pre-test to the post-test. This student, in grade 9, responded at Code 2 on the first two occasions and at Code 3 for the longitudinal follow-up.

- “Random means not in any particular order. Tattslotto is done randomly. So is pulling names out of a hat” [pre-test].
- “In no particular order. Weather” [post-test].
- “When something happens in a total unpatterned or unmethodolic way. Rolling a die 60 times. It will randomly land on the numbers” [longitudinal].

For the term “Variation” grade 7 students displayed greater changes in understanding the term than grade 9 students did. The following is an example of an increase from the pre-test (Code 0) to the post-test (Code 1) and a further increase from the post-test to the longitudinal follow-up (Code 2).

- “Don’t know” [pre-test].
- “(a) A variation of something like food etc. (b) The variation of food was amazing. (c) Food, shoes, events, shops” [post-test].
- “(a) Lots of different things. (b) The variation of chocolates that were on the bench. (c) Days in school (subjects)” [longitudinal].

The development of understanding of the term for this student was gradual and the final response did not show a complete grasp of the term. This was a common trend for students in grade 7 answering this question. The following student responded at a consistently higher level than the last student, however, once again highlighting the increase from the pre-test (Code 1) to the post-test (Code 2), to the longitudinal follow-up (Code 3).

- “(a) Something varies from the other. (b) The variation between a cat and a dog. (c) Cat and dog” [pre-test].
- “(a) It means something’s different. (b) I haven't got any more shoes in red but the variation is green. (c) Shoe colours” [post-test].
- “(a) Something that has changed. (b) The double kick variation of a skateboard is a lot better. (c) The stock market” [longitudinal].

The pre and longitudinal means and standard errors and the results of paired *t*-tests for each question for Sample 3, the group without specialised instruction, are given in Table 7. For “Sample” the greatest improvement occurred for students moving from grade 3 to 5 but improvement also occurred for students who were originally in grade 7. For the students who were originally in grade 7, improvement from the pre-test to the longitudinal follow-up also occurred for the terms “Random” and “Variation,” with the greatest improvement for “Variation.” The students who were originally in grade 9 did not improve significantly on any of the questions, and in fact showed decreasing performance on “Sample.”

Table 7
Pre and Longitudinal Means, Standard Errors and Paired t-tests for Sample 3

		Grade 3/5*	Grade 5/7*	Grade 7/9*	Grade 9/11*
Sample	Pre Mean, Std. Error	0.37, 0.033	0.91, 0.097	1.13, 0.118	1.65, 0.150
	Longitudinal Mean, Std. Error	0.90, 0.100	0.95, 0.118	1.43, 0.086	1.48, 0.142
	t, p	-4.88, p<.001	-0.31, p=.38	-2.61, p<.01	0.72, p=.24
Random	Pre Mean, Std. Error			0.85, 0.101	1.03, 0.138
	Longitudinal Mean, Std. Error			1.24, 0.093	1.23, 0.116
	t, p			-2.78, p<.01	-0.95, p=.18
Variation	Pre Mean, Std. Error			0.90, 0.123	1.29, 0.194
	Longitudinal Mean, Std. Error			1.52, 0.135	1.58, 0.165
	t, p			-4.86, p<.001	-1.22, p=.12

*Grade in Longitudinal Follow-up

Table 8 shows the means of difference scores for Sample 2B, which received specialised instruction, and for Sample 3, which did not receive specialised instruction, based on longitudinal follow-up scores compared to pre-test scores. The mean of the difference scores (longitudinal – pre score) were compared for each grade and question for Samples 2B and 3, with *t*-tests showing there were no differences between the two samples for any grade or on any question.

Table 8
Pre and Longitudinal Mean Differences, Standard Errors and t-tests for Samples 2B and 3

		Grade 3/5*	Grade 5/7*	Grade 7/9*	Grade 9/11*
Sample	S2B Mean Difference, Std. Error	0.36, 0.155	0.33, 0.139	0.55, 0.145	0.36, 0.269
	S3 Mean Difference, Std. Error	0.52, 0.094	0.05, 0.140	0.30, 0.107	-0.16, 0.291
	t, p	-0.90, p=.19	1.44, p<.08	1.44, p<.08	1.63, p<.06
Random	S2B Mean Difference, Std. Error			0.36, 0.121	0.36, 0.255
	S3 Mean Difference, Std. Error			0.39, 0.159	0.19, 0.233
	t, p			-0.16, p=.44	-0.01, p=.49
Variation	S2B Mean Difference, Std. Error			0.66, 0.135	0.29, 0.222
	S3 Mean Difference, Std. Error			0.63, 0.136	0.29, 0.314
	t, p			0.16, p=.44	-0.01, p=.49

*Grade in Longitudinal Follow-up

Research Question 4: Association across Terms

Table 9 displays the correlation coefficients between each pair of the questions for the students in Sample 1. As can be seen there is a reasonably strong association between the three terms (p<.01, for all pairs), although the amount of variation explained ranged from 19% to 23%.

Table 9
Correlations between the Terms "Sample," "Random," and "Variation"

	Random	Variation
Sample	0.433	0.481
Random		0.463

Discussion

The Discussion is presented in two parts. The first considers observations of interesting results in relation to the cohorts, to change over time, and to the influences of instruction on vocabulary.

The second part considers implications for those concerned about statistics, literacy, and statistical literacy.

Explanation of Results

Several aspects of the results warrant attention and an attempt at the explanation. In considering the initial performance of students in the four grades in Sample 1, the plateau in average score from grade 5 may indicate a lack of emphasis on literacy skills in describing mathematical ideas in the middle school, whereas the increase from grade 3 to 5 may be a natural increase in language skills. In relation to this lack of emphasis on literacy skills in relation to mathematical ideas, Malone and Miller (1993) and Miller (1993) found that teachers do not always use correct terminology in their instructional language, as they do not think students would be familiar with the terminology and instead use everyday language. With the recent interest in teaching statistics, some teachers may even be uncomfortable themselves with statistical terminology and therefore avoid the use of these terms in instruction to avoid having to explain the meaning to the class.

The overall improvement in performance after instruction for grades 3, 5, and 7 is encouraging and would appear to indicate that ideas related to "sample," "random," and "variation" were picked up from classroom activities. The lack of improvement for grade 9 cannot be explained in terms of a ceiling effect for performance. There is a possibility that it could be the result of a reduction of interest in the mathematics curriculum in general by that grade or perhaps a plateau in the development of literacy skills that would assist in expressing ideas related to the concepts in words.

A comparison of performance after two years between those who did and did not experience instruction provided by the research project, however, showed virtually no difference in performance between the four pairs of grades. It is likely that a combination of possibilities explains this outcome. It may be that involvement with the research team resulted in the schools who had had special instruction subsequently ignoring the chance and data curriculum and hence minimising further improvement. At the same time it is known that some teachers in the control schools attended Quality Teacher Program professional development and hence may have provided equivalent experiences over the two-year period.

Since the grade 9 students' performances were similar over the two groups it would appear that something more general than the effect of the research project's instruction is associated with the lack of improvement by grade 11. The much reduced sample size for the grade 9-11 cohort could cause concern in terms of the higher ability students being expected to continue to grade 11 (seen in the means for the subsets in Samples 2B and 3). The use of paired *t*-tests, however, takes this into account.

Overall the changes in performance, regardless of whether initiated by specialized instruction by a research project or initiated by teachers on their own, point to the importance of the middle school years in improving an appreciation of the concepts underlying the chance and data curriculum. It would appear the focus should be before grade 9 if activities are based on concrete experiences in handling data. It may be that a more theoretical approach will prove effective from grade 9 but this was not trialed in the current project. The experience of the authors in classrooms however, leads them to suspect that this is not the solution except for students in advanced mathematics streams.

The association of scores between the pairs of terms (see Table 9) may reflect a similar underlying appreciation of the terms or it may reflect an underlying literacy skill that assists students in expressing the concepts in words. That it is not stronger probably reflects the lack of explicit discussion in the classroom and lack of expectation for students to verbalise their understanding of the concepts. Further research linked to other aspects of literacy may be useful in explaining the relationship. The overall association of the three levels of performance (codes 1, 2, and 3) for each term with Levels 2, 3, and 4, respectively, of the construct identified with students' understanding of variation across the chance and data curriculum (Watson et al., 2003) suggests that both statistical and literacy understanding may be involved.

Implications

From an educational standpoint a depressing aspect of the data reported in this study is the high percent of idiosyncratic responses across grades for all three terms. By grade 9, 30% of students in Sample 1 were still not able to define "sample," 48% were unable to define "random," and 50% were unable to define "variation," describing the terms in an ikonic fashion. Some Level 0 responses may be accounted for by an unwillingness to answer a survey or an inability to express complex ideas in words. Results obtained by Malone and Miller (1993) and Miller (1993) reveal a similar result with a very small percentage of students being able to define mathematical terms in their own language without the use of symbols. For example for the term "factor," Malone and Miller found that only 8.5% of grade 8 students, 6% of grade 9 students, 4% of grade 10 students, and 7% of grade 11 students gave acceptable written responses.

Through using impromptu writing techniques, Miller (1993) found that students' use of mathematic vocabulary is very limited and there are instances where students confuse everyday words with less familiar mathematical terms, for example, "factor" and "factory." In the current study, similar occurrences with "sample" and "simple" and "random" and "ransom" were common. Students, however, seemed to demonstrate an even more limited understanding in the current study by giving an acceptable definition to the mistaken term. For example in response to the question on "sample," a grade 7 student responded "Not very hard. The cake was simple to make." Similarly in response to the question on "random" a grade 9 student responded with "When you want money. You highjack a plane and ask for money." These examples show a clear confusion between the traditional meaning of a word and an unfamiliar one that sounds similar. It would therefore appear that more practice is needed in the classroom to express ideas and concepts in words.

One question that occurs in relation to defining statistical terms is what model is set before students. Terms such as "random" and "variation" are often treated in textbooks and also by leaders in the field as undefined terms. David Moore's (1990) description of "random" shows what a difficult concept it is in a statistical context.

> Phenomena having uncertain individual outcomes but a regular pattern of outcomes in many repetitions are called *random*. 'Random' is not a synonym for 'haphazard,' but a description of a kind of order different from the deterministic one that is popularly associated with science and mathematics. Probability is the branch of mathematics that describes randomness. (p.98)

The fact no one in this study came close to the richness of Moore's description is not surprising. Probably many teachers would have a similar difficulty. When curriculum documents (e.g., New South Wales Board of Studies, 2002) contain outcomes such as "students learn to recognise randomness in chance situations" (p. 75) with no further description of randomness, one wonders where the foundation for understanding will come from.

Even David Moore (1990), who provides the excellent description of "random" above, spends a great deal of effort explaining the importance of "variation" as the foundation of statistics without defining the term. Many curriculum documents do not even use the word "variation" in relation to what happens in the chance and data curriculum. Instead the focus is often on average, as an elementary way to describe data sets. Without variation, however an average is an unnecessary tool. Terms like range, interquartile range, and standard deviation seek to describe variation in data sets but without explicit connections being made it is doubtful if all students appreciate the role of these statistics.

"Sample" is a somewhat easier idea than the other two and many books that discuss sampling provide an adequate description. The description of "sample" below is from Orr (1995).

> Sometimes, for theoretical, practical, or efficiency reasons, it is desirable to study (collect data from) less than the entire population. In such cases a subset of the population called a sample is selected. Although data are then collected only from or about the sample, conclusions are drawn (generalized) to the larger population as well … What is the essential nature of a sample? In a word, a sample should be a small-scale replica of the population from which it is selected, in all respects that might affect the study conclusions. (p. 72)

At the primary/middle school level books are inconsistent, with Corwin and Friel (1990) providing an excellent description of "Sample" for grades 5 and 6, whereas Bereska, Bolster, Bolster, and Schaeffer (1999) omit the term from their glossary for grades 4 to 8. What is potentially interesting in the case of "sample" is the fact that many students see "sample" in a homogeneous sense of representing exactly what some population, probably a product for sale, is like. In statistics, however, a sample is important as a heterogeneous representation of the variation present in the population. That variation may be the result of a random phenomenon or may be caused by some naturally explained phenomenon.

The linking of these three ideas necessary for higher order thinking indicates that the school curriculum needs to be more explicit in addressing fundamental literacy issues, starting in the middle school years with more explicit reference to concepts and examples. The fact that many students can provide examples of samples, random phenomena, and things that vary, indicates that there are starting points. Teachers across the curriculum often encounter topics using these three ideas and if made aware of the need, could stress the importance of the concepts and expect students to use them meaningfully in written work. Although not excusing the mathematics teachers from having similar expectations, such action would be excellent reinforcement. As states adopt revised curricula with "New Basics" (Education Queensland, 2000) and "Essential Learnings" (Department of Education Tasmania, 2002), there is both an opportunity and a responsibility to link ideas such as these across the topics in the "new" learning areas.

Conclusion

This study has raised as many questions as it has answered about students' understanding of vocabulary associated with statistical literacy. It appears that both specific instruction and other instruction over time can improve students' appreciation of the associated concepts over the middle years. In general this may be a recent innovation in some Tasmanian middle schools. The difficulty and complexity of describing the three terms is seen not only in the continuing moderate mean scores but also in the distribution of code levels across the levels of the overall construct associated with understanding of variation with regard to chance and data. More research related to other aspects of literacy is likely to be helpful in making suggestions for further improvement in performance.

Acknowledgements

This research was funded by an Australian Research Council grant (No.A00000716).

References

Bereska, C., Bolster, C. H., Bolster, L. C., & Schaeffer, R. (1999). *Exploring statistics in the elementary grades. Book two (grades 4-8)*. White Plains, NY: Dale Seymour.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Biggs, J. B., & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behaviour. In H. A. H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 57-76). Hillsdale, NJ: Lawrence Erlbaum.

Bridge, D. (2002). *Outcomes and outputs from literacy and numeracy programs, projects, trials and research 2000-2002: Summary review*. Hobart: Department of Education Tasmania.

Corwin, R. B., & Friel, S. N. (1990). *Statistics: Prediction and sampling: A unit of study for grades 5-6*. [Used numbers: Real data in the classroom] Palo Alto, CA: Dale Seymour Publications.

Department of Education Tasmania. (2000). *The Tasmanian literacy and numeracy plan 2000-2002*. Hobart: Author.

Department of Education Tasmania. (2002). *Essential learnings framework 1*. Hobart: Author.

Education Queensland. (2000). *New Basics Project technical paper*. Retrieved 28 May, 2003 from http://education.qld.gov.au/corporate/newbasics/html/library.html

Luke, A., & Freebody, P. (1997). Shaping the social practices of reading. In S. Musprati, A. Luke, & P. Freebody (Eds.), *Constructing critical literacies: Teaching and learning textual practice* (pp. 185-225). St. Leonards, NSW: Allen and Unwin.

Malone, J., & Miller, D. (1993). Communicating mathematical terms in writing: Some influential variables. In M. Stephens, A. Waywood, D. Clarke, & J. Izard (Eds.), *Communicating mathematics: Perspectives from classroom practice and current research* (pp. 177-190). Hawthorn, VIC: The Australian Council for Educational Research Ltd.

Miller, L. D. (1993). Making the connection with language. *Arithmetic Teacher, 40*, 311-316.

Moore, D. S. (1990). Uncertainty. In L. S. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.

Moritz, J. B., Watson, J. M., & Pereira-Mendoza, L. (1996, November). *The language of statistical understanding: An investigation in two countries*. A paper presented at the Joint ERA/AARE Conference, Singapore. http://www.swin.edu.au/aare/96pap/morij96.280

New South Wales Board of Studies. (2002). *Mathematics: Years 7-10 syllabus*. Sydney: Author.

Orr, D. B. (1995). *Fundamentals of applied statistics and surveys*. New York: Chapman & Hall.

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

Watson, J. M. (1994). Instruments to assess statistical concepts in the school curriculum. In National Organizing Committee (Ed.), *Proceedings of the Fourth International Conference on Teaching Statistics. Volume 1* (Vol. 1, pp. 73-80). Rabat, Morocco: National Institute of Statistics and Applied Economics.

Watson, J. M., Collis, K. F., & Moritz, J. B. (1993, September). *Assessment of statistical understanding in Australian schools*. A paper presented at the Statistics '93 conference, Wollongong, NSW.

Watson, J. M., & Kelly, B. A. (2002a). Can grade 3 students learn about variation? In B. Phillips (Ed.), *Developing a statistically literate society* (CD of the Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town). Voorburg, The Netherlands: International Statistical Institute.

Watson, J. M., & Kelly, B. A. (2002b). Variation as part of chance and data in grades 7 and 9. In B. Barton, K. C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds.), *Mathematics education in the South Pacific* (Proceedings of the 26th annual conference of the Mathematics Education Research Group of Australasia, Auckland, pp. 682-689). Sydney: MERGA.

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.

Watson, J. M., & Moritz, J. B. (2000). Development of understanding of sampling for statistical literacy. *Journal of Mathematical Behavior, 19*, 109-136.