Variation as Part of Chance and Data in Grades 7 and 9

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

Ben A. Kelly
University of Tasmania
<Ben.Kelly@utas.edu.au>

As part of a larger project studying school students' understanding of statistical variation, 92 students in grade 7 and 90 students in grade 9 participated in a unit of work related to the chance and data curriculum, emphasising variation with respect to the topics covered. Students completed pre- and post-tests devised to assess understanding before and after the lessons in the unit. This paper reports on the teaching arrangements for the classes taking part and the change in performance after the unit.

It appears to have taken a long time for educators to catch up with statisticians in recognising the importance of variation within the realms of statistical enquiry. As statistics was emerging as part of the school mathematics curriculum (National Council of Teachers of Mathematics [NCTM], 1989; Australian Education Council [AEC], 1991), it was David Moore (1990) who stressed the omnipresence of variability. Unfortunately explicit mention in relation to variation in the Australian *National Statement* was limited to "measures of spread" (e.g., AEC, p. 178). Although followed by calls from Green (1993) and Shaughnessy (1997) for research into students' understanding of variability, there is still little reported research on school students' understanding and learning of the topic. Following the initial research of Shaughnessy, Watson, Moritz, and Reading (1999), based on drawing lollies from a container of 100 with 50 red, Reading and Shaughnessy (2000) and Torok and Watson (2000) suggested developmental sequences for the understanding of variation throughout the years of schooling.

The data reported here were collected as part of a larger follow-up study after the work of Shaughnessy et al. (1999), in particular to document change, if it occurred, resulting from a planned classroom intervention program focussing on variation within chance and data. Students in grades 3, 5, 7, and 9 participated, with different teaching plans employed at the primary (grades 3 and 5) and secondary (grades 7 and 9) levels. This paper will report on interventions planned and outcomes achieved at the secondary level.

Instructional Design

Plans for teaching units for the chance and data curriculum were devised based on an analysis of the content as suggested by organisations such as the NCTM (1989) and the AEC (1991), and respected statistics educators such as Holmes (1980) and Moore (1990). Taking into account the process of the statistical analysis – data collection and sampling, data representation, data reduction, inference, and probability – aspects of variation associated with each were identified for focus lessons.

Although for the primary grades the teacher was provided from the project team and all lessons were nearly identical in content and length, at the secondary level it was necessary to fit in with the schools' timetables and rely on the usual mathematics teachers to deliver the units of work. In light of this an extensive package was prepared covering six possible units of work emphasising variation within the topics covered in the chance and data curriculum. Some units were based upon published materials, whereas others were designed by the first author. They included variation with spinners, with dice, in sampling,

B. Barton, K. C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds.) *Mathematics Education in the South Pacific* (Proceedings of the 25th annual conference of the Mathematics Education Research Group of Australasia, Auckland, pp. 682-689). Sydney: MERGA.

with associations, in comparing two groups, and in numbers of chocolate chips in cookies. Each unit could be adapted for between one and three lessons. The basic topics were chosen from those in the chance and data curriculum for grade 7 or 9. The unit on variation in associations, for example, dealt with the relationship of hand span to foot length, and how variation in the data affects the conclusions reached. Teachers received 21 pages of plans and 33 pages of associated documents (e.g., work sheets, copies from relevant books).

The authors met with the teachers involved in the project to explain the purpose and distribute the pages of plans and associated documents for the units. Suggestions were made as to the order in which units might be taught but the decisions on which to choose were left to the professional judgement of the teachers. It was realised that the units might take more time to cover comprehensively than was available in the teachers' programs. None of the teachers, however, had covered chance and data yet for the year in question.

As it turned out, the greatest variation to occur in the project overall was the variation in the amount of time spent on chance and data by the secondary teachers. Teachers could not be held to an early commitment that would have made the number of lessons in the secondary schools the same as in the primary schools, or even the same as each other. It was therefore necessary to include another variable in the study: "class".

Methodology

Sample. The students who participated in the project were from two high schools, one suburban and one semi-rural in the Australian state of Tasmania. In one school, two grade 7 classes ($n_1 = 16$ and $n_2 = 21$) were taught by the same teacher, and two grade 9 classes ($n_3 = 11$ and $n_4 = 15$) were taught by different teachers. In the other school, all classes were taught by different teachers, three in grade 7 ($n_5 = 14$, $n_6 = 14$, and $n_7 = 27$) and three in grade 9 ($n_8 = 31$, $n_9 = 16$, and $n_{10} = 17$). Not all students completed both pre- and post-tests and the sample sizes reported here reflect the number of students with complete data, which may be a smaller number than the number present when teaching occurred. In grade 7 there were 34 girls and 58 boys and in grade 9 there were 39 girls and 51 boys.

Survey instruments. The survey instruments for the two grades were the same except for one additional item used with grade 9 students. All items are presented in Watson, Kelly, Callingham, and Shaughnessy (2001) and were clustered into four subscales measuring understanding of Basic Chance and Data (BCD), Chance Variation (CV), Data Variation (DV), and Sampling Variation (SV). Items in the BCD subscale included evaluating probabilities associated with random generators and drawing objects from boxes, and reading information from a pictograph and a two-way table. Items on CV asked for predictions of 60 outcomes of tossing a die, predictions of repeated trials with a 50-50 spinner, and a definition of the term random. DV questions involved inferences from graphs of different visual appearance, identifying false or unusual data from different graphs, defining variation, and calculating the average from a data set containing an outlier. Items on SV were based on a classroom scenario where a survey is framed and a list of possible sampling methods presented for evaluation, the definition of a sample, two fair ways of selecting a sample, and a news article stating evidence based on a biased sample. It was hence possible to measure change after the teaching, on the four subscales as well as the entire test. The post-test in each grade was identical to the pre-test. The tests were administered by the teachers at convenient class times before and after instruction within time windows requested by the research team.

Analysis. Responses were entered into a spreadsheet and after agreement of the authors on expected levels of response, coding was carried out by the second author. A clustering of similar answers took place on the spreadsheet and both authors checked the consistency of coding, with discrepancies resolved by discussion. Paired *t*-tests were performed for each class and for each grade overall, for the four subscales and the total of all items. Within each grade, a one-way ANOVA was performed for "post-test score – pre-test score", with class as a factor. The gender imbalance of the two grades (37% girls in grade 7 and 43% girls in grade 9) suggested a further comparison of scores by gender.

Results

The results will be discussed in four parts: outcomes for individual classes and grades, differences across classes, examples of responses to particular items, and gender differences. The pre- and post-test means, as well as standard errors of the differences, for each of the five grade 7 classes on the four subscales and the total scale are given in Table 1. Similar results for the five grade 9 classes are given in Table 2. The total possible scores for each of the common subscales for the two grades were 22 for BCD, 19 for CV, and 32 for DV. For SV, in grade 7 the total was 30 and in grade 9 it was 33. The Total score in grade 7 was 100 and in grade 9 it was 103. The item asking for a definition and example of "variation" was used in both the CV and DV subscales, but only counted once in the Total.

Table 1
Pre- and Post-test Means and Standard Errors of the Differences for Grade 7

Class Scale	7A (n = 16)	7B (n = 21)	7C (n = 14)	7D (n = 14)	7E (n = 27)	Total (n = 92)
BCD	15.3 / 17.1 (0.70)*	12.9 / 17.3 (0.86)***	10.9 / 12.7 (0.65)**	11.4 / 9.6 (1.34)	12.6 / 16.4 (0.73)***	12.7 / 15.1 (0.44)***
CV	9.1 / 11.1 (0.73)**	8.4 / 12.2 (0.39)***	6.3 / 8.8 (0.42)***	7.0 / 6.9 (0.96)	7.6 / 10.2 (0.61)***	7.8 / 10.1 (0.31)***
DV	11.9 / 15.4 (1.31)**	11.8 / 17.1 (0.90)***	9.1 / 12.9 (1.06)**	10.1 / 8.1 (0.99)*	9.9 / 15.1 (0.72)***	10.6 / 14.2 (0.50)***
SV	9.6 / 14.7 (1.09)***	10.0 / 16.1 (1.09)***	4.6 / 7.7 (1.12)**	7.5 / 3.7 (1.26)**	7.1 / 11.5 (1.15)***	7.9 / 11.3 (0.62)***
Total	44.9 / 56.7 (2.80)***	42.3 / 61.3 (1.96)***	30.5 / 41.7 (1.42)***	35.8 / 28.3 (2.63)**	36.5 / 51.9 (2.06)***	38.3 / 49.7 (1.33)***

*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Outcomes for individual classes and grades. Paired *t*-tests for the combined samples for each of grade 7 and grade 9 indicated that for each grade, there was a significant improvement in the Total score ($p < 0.001$). For the classes in this study, the pre-test mean for grade 7 was significantly less ($p < 0.01$) than that for grade 9 but improvement was greater resulting in a post-test mean for grade 7 very close to that for grade 9. For the five grade 7 classes, differences represented significant improvement for four classes ($p < 0.001$) and a diminished performance for one class ($p < 0.05$). For grade 9, improvement occurred for four classes (two at the 0.001-level, one at the 0.05-level, and one non-significantly), whereas one class showed a diminished performance ($p < 0.001$). Classes 7D and 9I showed a decrease in performance in nearly all subscales, as well as overall.

Table 2
Pre- and Post-test Means and Standard Errors of the Differences for Grade 9

Class Scale	9F (n = 11)	9G (n = 15)	9H (n = 31)	9I (n = 16)	9J (n = 17)	Total (n = 90)
BCD	14.5 / 14.6 (0.92)	9.3 / 15.3 (1.10)***	14.3 / 17.2 (0.78)***	14.2 / 10.8 (1.21)**	11.1 / 11.1 (1.12)	12.9 / 14.3 (0.55)**
CV	10.5 / 10.5 (0.86)	7.2 / 11.1 (0.85)***	9.4 / 11.7 (0.63)***	9.2 / 6.9 (0.96)*	7.1 / 8.8 (0.70)*	8.7 / 10.1 (0.40)***
DV	13.8 / 15.9 (1.34)	7.7 / 14.6 (0.98)***	14.4 / 17.4 (0.68)***	12.3 / 8.4 (1.20)**	7.6 / 9.8 (1.13)*	11.6 / 13.7 (0.55)***
SV	14.0 / 16.7 (1.10)*	6.5 / 14.4 (1.59)***	13.5 / 17.3 (1.05)***	12.6 / 4.9 (1.24)***	8.5 / 10.2 (1.78)	11.3 / 13.2 (0.79)**
Total	51.0 / 56.3 (2.42)*	30.1 / 54.1 (2.72)***	50.8 / 62.1 (2.04)***	47.1 / 30.9 (3.19)***	34.1 / 39.4 (3.45)	43.6 / 50.2 (1.78)***

*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

In each grade the BCD scale post-test mean was the greatest of the subscales in relation to the total possible score (69% for grade 7 and 65% for grade 9). All other subscale post-means were similar relative to the totals possible and 53% of the total or less.

For all four subscales and the Total, the average improvement for grade 9 on common items was less than for grade 7. Although in each case the grade 9 pre-test average was higher, in two cases the grade 7 post-test average was slightly higher. With the highest class average 62.1 out of 103, it is unlikely a ceiling effect was operating for the grade 9 classes. The greatest variation in subscale scores occurred for the SV subscale for both grades.

Differences across classes. Although it is clear from the tables that the improvement in Total scores varies greatly between classes at each grade, one-way ANOVAs on "post-test score – pre-test score" confirms that differences are significant at each grade level ($p < 0.001$). The "class" variable hence was associated with significant differences in "post-test – pre-test" change within the two grades.

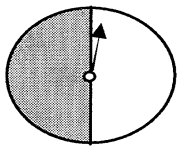
Based on the journals provided by the teachers, estimates were made for the number of lessons taught in each of the classes. Table 3 summarises the "post – pre" differences in Total means for the classes and the number of lessons taught, ordered by the number of lessons taught. It is clear that the association of the number of lessons taught and change in Total mean score from pre- to post-test is not linear, or even particularly strong. With $r = 0.428$ and 18.3% of variance explained, the correlation is not significant ($p = 0.10$). The conclusion is reached that other factors besides the number of lessons taught are associated with the differences in "post – pre" Total means.

Examples of responses to particular items. Paired *t*-tests for each item by grade level revealed many items in which the grade 7 students improved significantly from the pre- to post-test for which the grade 9 students did not. On particular items, like the first three parts of the Spinner item in Figure 1, the grade 9 students' mean pre-test score was higher than the grade 7 mean pre-test score. This indicates that there was more room for improvement for the grade 7 students. On other items, like Part 4 of the Spinner item in Figure 1, grade 7 students had a slightly higher pre-test mean and improved more than the grade 9 students.

Table 3
Number of Lessons Taught per Class with Post – Pre Difference Means

Class	7C	9H	7D	9I	9F	7A	7B	9G	7E	9J
Number of Lessons	3	3	5	5	6	9	9	12	14	14
Post – Pre Difference Mean	10.9	11.1	-6.3	-15.5	5.2	10.5	19.1	23.8	14.8	5.2

A class used this spinner.



1. Out of 50 spins, how many times do you think the spinner will land on the shaded part? Why do you think this?
2. If you were to spin it 50 times again, would you expect to get the same number out of 50 to land on the shaded part next time? Why do you think this?
3. How many times out of 50 spins, landing on the shaded part would surprise you?
4. Suppose that you were to do 6 sets of 50 spins. Write a list that would describe what might happen for the number of times the spinner would land on the shaded part?

_____, _____, _____, _____, _____, _____

Figure 1. Spinner item used on pre-and post-test.

Figure 2 shows several responses from grades 7 and 9 to the four parts of the Spinner task in Figure 1. The ranges of codes for these items were 0-3, 0-3, 0-1, and 0-2, respectively, and the pre- and post-means are given for each grade. Responses are typical of the coding levels that contributed to the means and to some degree to the improvement or lack of it in the post-test scores. To obtain the highest code (3) on Parts 1 and 2 some notion of variability or uncertainty was required. Students who acknowledged strict probability with no variation received a code of 2, whereas students who gave imaginative answers were given a code of 1. For Part 4, criteria were set for the variation displayed. Responses that did not involve numbers or were outside the 0-50 range were coded 0, those with no variation or too much variation were coded 1, and those with reasonable variation were coded 2. It was difficult for many students to suggest reasonable variation.

Gender differences. For grade 7, although the pre-test average for girls was higher on four of the five scales ($p < 0.03$), the average "post – pre" score for girls was not significantly different from the boys on any scale. The mean improvement for girls was 10.8 and for boys it was 11.8. Only for the DV scale did the girls continue to have a higher average score than boys on the post-test ($p < 0.04$). In most cases, for subscale scores and differences, the variation displayed in boys' scores was greater than in the girls' scores.

The relationship of scores for girls and boys in grade 9 was somewhat different from grade 7. The girls' average pre-test scores were significantly higher on only two subscales ($p < 0.05$) but were significantly higher on all post-test scores ($p < 0.03$ to $p < 0.001$). Although the "post – pre" score average was higher for girls on each scale, the difference was significant on only the BCD scale ($p < 0.01$) where the boys' average did not change. Again the variances in boys' difference scores were generally greater than for girls. Despite indications of better performance in various subscales on the pre- or post-tests by girls, overall the improvement was not significantly different in either grade.

1. Out of 50 spins, how many times do you think the spinner will land on the shaded part? Why?			
Grade 7 pre-mean = 1.91, post-mean = 2.16		Grade 9 pre-mean = 2.00, post-mean = 1.96	
PRE: 40, Because the arrow needs energy to come up [code 1]	POST: 25, Because the chance to land on the shaded and white part are the same [code 2]	PRE: 25, Because there is an equal chance [code 2]	POST: 25, Because there's equal chances of it landing on each part [code 2]
PRE: 29, Because that one's better [code 1]	POST: 20-25, there's only 2 parts and there's still a 50/50 chance [code 3]	PRE: 23, Because there is a 50-50 chance but some sides get hit a bit more [code 3]	POST: 19, Because I like white [code 1]
2. If you were to spin it 50 times again, would you expect to get the same number out of 50 to land on the shaded part next time? Why do you think this?			
Grade 7 pre-mean = 1.42, post-mean = 1.70		Grade 9 pre-mean = 1.67, post-mean = 1.90	
PRE: No, because it might have been pushed harder … it might land a couple of numbers different [code 1]	POST: No, because the force of your spin might change the side it goes on [code 1]	PRE: Probably not, because the saying goes – lightening only strikes once [code 1]	POST: Probably not, there are many different things that could occur, like how hard you spin it … [code 1]
PRE: No, … you might spin it harder or softer at any time [code 1]	POST: No, it might be but it would be a tiny difference [code 3]	PRE: No, because it matters how hard you spin it [code 1]	POST: No, because it's all up to chance [code 2]
3. How many times out of 50 spins, landing on the shaded part would surprise you?			
Grade 7 pre-mean = 0.65, post-mean = 0.84		Grade 9 pre-mean = 0.77, post-mean = 0.83	
PRE: Half [code 0]	POST: 50 [code 1]	PRE: Luck [code 0]	POST: None [code 1]
PRE: All of them [code 1]	POST: 50 [code 1]	PRE: Below 10 or above 45 [code 1]	POST: Anything above 40; 50 would really surprise me [code 1]
4. Suppose you were to do 6 sets of 50 spins. Write a list that would describe what might happen for the number of times the spinner would land on the shaded part?			
Grade 7 pre-mean = 1.04, post-mean = 1.43		Grade 9 pre-mean = 0.92, post-mean = 1.22	
PRE: 25, 25, 25, 25, 25, 25 [code 1]	POST: 25, 25, 25, 25, 25, 25 [code 1]	PRE: ?, ?, ?, ?, ?, ? [code 0]	POST: 25, 15, 30, 39, 28, 42 [code 1]
PRE: Shade, shade, black, shade, shade, black [code 0]	POST: 23, 21, 22, 26, 20, 24 [code 2]	PRE: 25, 25, 25, 25, 25, 25 [code 1]	POST: 45, 10, 5, 20, 50, 0 [code 1]
PRE: 24, 26, 25, 24, 23, 24 [code 1]	POST: 27, 23, 25, 26, 22, 24 [code 2]	PRE: 24, 25, 25, 26, 25, 25 [code 1]	POST: 25, 28, 30, 24, 20, 18 [code 2]

Figure 2. Typical pre-and post-test responses to the Spinner item in Figure 1.

Discussion

In terms of overall performance of students, the significant gains give support to the view that it is possible to teach chance and data with an emphasis on variation and expect positive results. The fact that the post-mean was a higher percentage in relation to possible score for the BCD subscale compared to the others, could be explained in several ways. It may be that BCD ideas, being prerequisites for dealing with variation, were covered more thoroughly in previous years, or indeed by the teachers of the classes in this study. It could also be that, having less variable or uncertain answers, BCD questions are easier for

students who are used to unique answers in their study of mathematics. The other three subscales, however, all showed significant improvement overall for each grade, indicating that progress was made on integrating variation concepts with topics in chance and data.

The setting for the study in two schools that represented typical state school learning environments introduced several factors deserving attention in considering the outcomes. These will be considered in turn, in relation to the grades, classes, and number of lessons taught; optimal possible performance; potential gender differences; and teacher participation. The differences in pre-test averages for the grade 7 classes reflect school differences in that no streaming took place in either school at this grade. Differences in pre-test averages for the grade 9 classes do reflect streaming of students in the two schools, with classes 9F, 9G, and 9J considered to cater for lower ability students. This study was interested in change relative to starting point and not the starting point per se.

The number of lessons taught was in itself not an accurate predictor of the average improvement in test scores. Teacher 7E achieved excellent results from 14 lessons, whereas teacher 9J achieved positive but non-significant improvement from the same number of lessons. Noting, however, that the pre-test average for class 9J was lower than that for class 7E, it might be speculated that the concepts were more difficult for the students in class 9J and hence were not retained as well. It could also be that teacher 9J had to spend more time on ideas and that less progress was made overall. The fact that two classes (7D and 9I) had lower average scores in the post-test than in the pre-test is likely to reflect student attitude to completing the post-test, rather than an actual decrease in understanding. Students may have felt it unnecessary to provide details that they had earlier, or they may have had a negative reaction to being asked to complete the test again. The range of scores on the post-test Total was 1 to 81. This is yet another indication that students may not have taken the task seriously. The second post-test comment on Part 1 of the Spinner item from a grade 9 student ("I like white," see Figure 2), in the light of the high level pre-test comment, may in fact be intended to be facetious. If this happened often, it could have contributed to the lower post-test means for some classes.

Classes 7A and 7B with similar pre-test means, were taught virtually the same lessons, illustrating the variation in change for two similar conditions. The teacher reported anecdotally that class 7B was an eager class, willing to learn, whereas 7A was a struggle to teach. This appears to be reflected in the improvement shown.

Although significant improvements occurred for many classes, for many items, and for most subscales, performance on the post-test was still not optimal in terms of the expectations of the curriculum. In comparing change in grade 7 to change in grade 9 the question of a ceiling effect for grade 9 may arise. Although for some items (e.g., Part 3 of the Spinner item) the grade 9 pre-test mean was significantly higher than the grade 7 pre-test mean, it is difficult to claim that a ceiling effect was occurring for grade 9 students; in fact many times the mean for grade 9 did not reach 2/3 of the highest possible score for the item. If a ceiling effect operated, it reflected, on average, a less than optimal performance as judged by the authors in light of the desired outcomes.

Although the girls performed better on all pre- and post-scales, they did not appear to gain more from the lessons as judged by the "post – pre" difference in scores. In fact the boys' improvement was marginally better in grade 7. Hence it cannot be said that the lessons favoured one gender of student as gauged by this measure.

Finally, the teaching arrangements made in this study have an affinity with those of Helme and Stacey (2000) who provided resources to four teachers in a primary school with

the aim of improving student understanding of decimals. Despite expressing willingness to be involved, two teachers did not use the resources at all, one teacher used them "at least once", and the fourth teacher used seven of the activities. Helme and Stacey found that student improvement was strongly related to the teachers' use of the resources, a result more clear-cut than in the current study. It does confirm, however, the difficulty of holding teachers to commitments made before studies begin. Conducting research in actual school environments is fraught with such difficulties unless the project provides the teachers (as was done elsewhere in the larger project of which this study was a part). This is not to say that research such as the current work with grade 7 and 9 teachers or Helme and Stacey's project should not take place. The actual "real-world" school setting is where curriculum implementation will take place after research is finished and it is useful to have an expectation of how successful it is likely to be.

Acknowledgments

This research was funded by an Australian Research Council grant (No. A00000716). The authors thank Dr David Ratkowsky for helpful comments.

References

Australian Education Council. (1991). *A national statement on mathematics for Australian schools.* Carlton, VIC: Author.

Green, D. (1993). Data analysis: What research do we need? In L. Pereira-Mendoza (Ed.), *Introducing data analysis in the schools: Who should teach it and how?* (pp. 219-239). Voorburg, The Netherlands: International Statistical Institute.

Helme, S., & Stacey, K. (2000). Improving decimal understanding: Can targeted resources make a difference? In J. Bana & A. Chapman (Eds.), *Mathematics education beyond 2000* (Proceedings of the 23rd annual conference of the Mathematics Education Research Group of Australasia, pp. 299-306). Perth, WA: MERGA.

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

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

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics.* Reston VA: Author.

Reading, C., & Shaughnessy, M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89-96). Hiroshima, Japan: Hiroshima University.

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddilph & K. Carr (Eds.), *People in mathematics education* (Proceedings of the 20th annual conference of the Mathematics Education Research Group of Australasia, pp. 6-12). Waikato, NZ: MERGA.

Shaughnessy, J. M., Watson, J., Moritz, J., & Reading C. (1999, April). *School mathematics students' acknowledgment of statistical variation.* In C. Maher (Chair), There's more to life than centers. Presession Research Symposium conducted at the 77th annual National Council of Teachers of Mathematics conference, San Francisco, CA.

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal, 12,* 147-169.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2001). *The measurement of school students' understanding of statistical variation.* Manuscript submitted for publication.