

STATISTICAL VARIATION IN A CHANCE SETTING

Jane M. Watson and Ben A. Kelly
University of Tasmania, Hobart, Australia

A series of 13 survey questions based on a 50-50 spinner is used to explore school students' understanding of statistical variation in a chance setting. Five questions set the context by assessing understanding of theoretical expectation and representation of repeated trials in a stacked dot (line) plot. Four questions provide opportunity to display appreciation of variation from point expectation and four questions address variation from distributional expectation. A total of 707 students in grades 3 to 9 answered some or all of these questions. A subset of 334 students then took part in a unit of study on chance and data emphasising variation. These students answered a post-test including the same items. Analysis showed a progression across the years of schooling, plateauing at grade 7 and improvement for all grades after instruction. Implications are considered.

INTRODUCTION

Although research into school students' understanding of variation has not progressed as rapidly as research into other topics in the chance and data curriculum, interest in developing tasks that will allow students to display their appreciation of variation is growing. Zawojewski and Shaughnessy (2000) became frustrated with a national survey test item based on drawing 10 objects from a container with 50% red that required students to predict a specific number of reds. The item did not encourage a range of answers and Shaughnessy, Watson, Moritz, and Reading (1999) revised the item, employing various formats to allow students to describe not only the most likely outcome but also what variation might occur.

In a chance setting there are two aspects of variation that effect the distribution of the outcomes that appear over repeated trials. First is the variation from the theoretical value expected, the most likely outcome; for example, from a container with 50% red objects sometimes one might draw 4 red out of 10 or 6 red out of 10. This is the variation that produces the probability distributions, such as the binomial, which are studied in the senior secondary years. In practice an experimental distribution does not match exactly the theoretical distribution: there is also variation from the ideal shape of many outcomes. The development of tasks to reflect student understanding should give students opportunity to show appreciation of expectation (theory), of the likelihood of some variation from that expectation, of a distribution of expected outcomes, and of potential unrealistic variation from the distribution of expected variation.

Working in a straight-forward chance setting, as done by Shaughnessy et al. (1999), would appear to be important in order to focus on aspects of variation rather than details associated with working out probabilities or searching for causes of change. To explore appreciation of expected variation in a chance setting, students can be asked to predict a number of individual outcomes or to draw a distribution of outcomes. Both of these alternatives were tried with an interview protocol developed from the work of Shaughnessy et al. (1999) by Kelly and Watson (2002). Students had more difficulty with drawing a distribution than with predicting a list of possible outcomes. Even for students who could create a graph, the degree of variation displayed was usually greater than

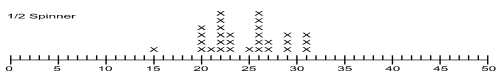
reasonable. The limited practical experience of students, however, suggested that it was unrealistic to expect a good fit to a theoretical model for a large number of trials.

Set 1. A class used this spinner.



- Q1. If you were to spin it once, what is the chance that it will land on the shaded part?
- Q2. Out of 10 (50) spins, how many times do you think the spinner will land on the shaded part? Why do you think this?
- Q3. If you were to spin it 10 (50) times again, would you expect to get the same number out of 10 (50) to land on the shaded part next time? Why do you think this?
- Q4. How many times out of 10 (50) spins, landing on the shaded part would surprise you?
- Q5. Suppose that you were to do 6 sets of 10 (50) spins. Write a list that would describe what might happen for the number of times the spinner would land on the shaded part?

Set 2. A class did 50 spins of the above spinner many times and the results for the number of times it landed on the shaded part are recorded below.



- Q6. What is the lowest value?
- Q7. What is the highest value?
- Q8. What is the range?
- Q9. What is the mode?
- Q10. How would you describe the shape of the graph?

Set 3. Imagine that three other classes produced graphs for the spinner. In some cases, the results were just made up without actually doing the experiment.

Q11. Do you think class A's results are made up or really from the experiment?



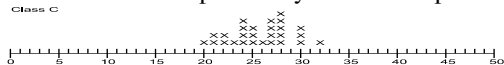
Made up Real from experiment - Explain why you think this.

Q12. Do you think class B's results are made up or really from the experiment?



Made up Real from experiment - Explain why you think this.

Q13. Do you think class C's results are made up or really from the experiment?



Made up Real from experiment - Explain why you think this.

Figure 1: Spinner questions used on survey

An alternative to asking students to produce their own distribution of outcomes is to present them with completed graphs representing repeated outcomes and ask which are reasonable and which are not. To do this requires a simple presentation, which like the basic chance setting, does not complicate the issue at hand. The stacked dot (or line) plot is well-suited to this task and has been found useful in various contexts for allowing students to focus on variation (Konold & Higgins, 2002), and on telling a story with data (Watson & Kelly, 2002c). The plot directly displays frequencies vertically as dots or Xs along a baseline with the scaled data values labeled.

This study is based on a series of 13 questions, shown in Figure 1. The scenario is the repeated spinning of a 50-50 spinner with the first set of five questions introducing the spinner and the potential for variation in outcomes in repeated trials. The second set of five questions introduces the stacked dot plot and explores student familiarity with the

basic characteristics of the graph of a distribution of outcomes in sets of repeated trials. The third setting offers three stacked dot plots, two with unlikely variation, e.g., too perfect or too much variation, and one with likely variation, for evaluation. The tasks are based on classroom activities developed by Torok (2000).

The research questions for this study are based on the usefulness of the tasks to assess student appreciation of variation in a chance setting and change after instruction. For a subscale of five Prerequisite questions (PRE), Q1, Q6, Q7, Q8, and Q9: Do students appreciate the theoretical chance involved and can they interpret a representation of a distribution? Is there a trend over grades and a difference after instruction? For a Point Estimate Variation subscale of four questions (PEVar), Q2, Q3, Q4, Q5: Do students' acknowledge the role of variation when predicting outcomes for a single and/or repeated trial of a spinner? Is there a trend over grades and a difference after instruction? For a Distributional Variation subscale of four questions (DisVar), Q10, Q11, Q12, Q13: How do students describe appropriate variation and can they accurately identify appropriate and inappropriate variation in an established distribution? Is there a trend over grades and a difference after instruction.

METHODOLOGY

Sample. Sample 1 consisted of 707 students from grades 3, 5, 7, and 9 in nine public schools in the Australian state of Tasmania who were surveyed as part of a larger study on school students' understanding of statistical variation. Sample 2 was a subset of 334 students from Sample 1 who received instruction on chance and data focusing on variation in five of the schools. These students were given a post-test to measure the effect of instruction on change. Sample sizes in the four grades are given in Table 1.

	Grade 3	Grade 5	Grade 7	Grade 9	Total
Sample 1	150	181	184	192	707
Sample 2	72	82	91	89	334

Table 1: Number of students per sample in each grade

Procedure. The questions in Figure 1 were part of a larger survey investigating students' understanding of statistical variation (Watson & Kelly, 2002a, 2002b). Students in grades 3 and 5 were presented with the Q1 to Q5 only and were asked Q2 to Q5 out of 10 spins. Students in grades 7 and 9 were given all thirteen questions, and were asked Q2 to Q5 out of 50 spins. All students were administered the surveys in class time. For grades 3 and 5 the researchers showed a spinner to the class before commencing the survey demonstrating the purpose of the spinner and how it worked in case there were any classes that may not have used spinners yet in their studies.

Students in Sample 2 were in the experimental group undertaking a unit of work on chance and data. Grades 3 and 5 were taught by the same teacher (provided by the research team), whereas grades 7 and 9 were taught by their usual mathematics teacher. Details of the lessons taught can be found in Watson and Kelly (2002a) for grades 3 and 5, and Watson and Kelly (2002b) for grades 7 and 9. During the intervention, all classes in Sample 2 were exposed to lessons using spinners and all grade 7 and 9 classes had graphing activities, most involving stacked dot plots. The students in Sample 2 were re-administered the same survey approximately 6 weeks after the completion of the lessons, in the same way it was administered for Sample 1.

Analysis. For this study the prerequisite questions, Q1, Q6, Q7, Q8, and Q9 were coded on a correct-incorrect basis. For these questions a code of 1 was given to responses of “50,” “Half,” “50/50” or the equivalent for Q1, “15” for Q6, “31” for Q7, “15-31” or “16” for Q8, and “22 and/or 26” for Q9. Responses to all other questions were categorized and coded by the authors to reflect an increasingly sophisticated appreciation of variation. In particular, the criterion for determining the appropriateness of the variation displayed in responses to Q5 was based on a simulation of 1000 outcomes using an EXCEL spreadsheet. The standard deviation for each simulation was calculated and then plotted, and appropriate variation was determined by values within the middle 90% of the distribution (0.6 – 2.3 for 10 spins; 1.3 – 5.0 for 50 spins). Examples of responses for codes for Q2 to Q5 are given in Table 4 and examples for Q10 to Q13 are given in Table 5 of the Results section. The scoring rubrics developed for these items were devised specifically to reward an appreciation of variation. It is acknowledged that others might devise different rubrics for different purposes, for example for Q10.

Although F-tests were performed for the PEVar scale across grades, t-tests are reported here as appropriate to describe observed differences for pairs of grades. T-tests were used to compare grades 7 and 9 on the PRE and DisVar scales also. Paired t-tests were used for Sample 2 with respect to pre- and post-test scores. The results will be discussed in the order of the research questions, with respect to the PRE questions, the PEVar questions, and the DisVar questions.

RESULTS

Prerequisite Ideas

Questions Q1, Q6 to Q9, in Figure 1 do not address variation but are necessary or potentially relevant to the other questions. Table 2 contains the percent correct on each item for grades 7 and 9 in Sample 1. For Q1, the grade 3 percent correct was 22.0% and for grade 5 it was 62.4%.

	Q1	Q6	Q7	Q8	Q9
G7	80.4	76.1	34.2	27.7	2.2
G9	85.4	71.9	33.9	31.8	16.1

Table 2: Percent correct for PRE questions in Sample 1

For Sample 2, Table 3 contains the pre- and post- performance on the PRE questions. With the exception for Q6, there was an improvement on all questions for both grades 7 and 9. For grade 3 the corresponding change for Q1 was from 23.6% to 40.3%, and for grade 5 it was 62.2% to 68.3%. Overall instruction assisted the basic ideas.

		Q1	Q6	Q7	Q8	Q9
G7	Pre	83.5	81.3	34.1	31.9	2.2
	Post	89.0	80.2	50.5	50.5	28.6
G9	Pre	86.5	76.4	42.7	34.8	24.7
	Post	94.4	77.5	53.9	41.6	31.5

Table 3: Percent correct for PRE questions for Sample 2 before and after instruction

Point Estimate Variation

Table 4 contains examples of responses to the PEVar questions. Without acknowledging variation, students were more likely to give code 2 responses to Q2. Only 5.3% of grade 3 students acknowledged variation in Q2, increasing to 7.2% in grade 5 and 19.6% in

grade 7, but dropping to 12.5% in grade 9. For Q3, again appreciation of variation increased slightly with grade.

Code	Summary	Examples
Q2. Out of 10 (50) spins, how many times do you think the spinner will land on the shaded part? Why do you think this?		
3	Variation	24 because it is close to half; About 5, it's equal
2	Theoretically correct; Anything can happen	25 because there's an equal chance for shaded and white 25, divide by 2; 50/50, you can't estimate
1	Inappropriate reasoning	5 out of 10 because you have a good chance of getting this one; Very good ... it might land on the white side
0	NR	Don't know
Q3. If you were to spin it 10 (50) times again, would you expect to get the same number out of 10 (50) to land on the shaded part next time? Why do you think this?		
3	Variation	Not exactly because the spinner would vary slightly
2	Anything can happen; Chance	No, it's the luck of the spin; Yes, same odds
1	Non- chance theories	No, because nothing has changed at all Yes, because you do the same as you did the first time
0	NR/no reason	Yes, just guessing
Q4. How many times out of 10 (50) spins, landing on the shaded part, would surprise you?		
1	Reasonable responses	Grade 3 & 5 – 0,1,2,8,9,10; Grade 7 & 9 – <20, >30
0	Inappropriate responses	Grade 3 & 5 – 3,4,5,6,7, Grade 7 & 9 –20 to 30 Ambiguous (e.g., Not many), misinterpretation
Q5. Suppose that you were to do 6 sets of 10 (50) spins. Write a list that would describe what might happen for the number of times the spinner would land on the shaded part?		
3	Appropriate variation	5, 2, 6, 7, 8, 9 (SD = 0.6-2.3 – 10 spins) 30, 20, 32, 26, 24, 18 (SD = 1.3-5.0 – 50 spins)
2	Too small, too wide or no variation (strict probability)	5, 8, 10, 1, 3, 2, 1 (SD <0.6 >2.3 – 10 spins), 5, 30, 26, 18, 45, 50 (SD <1.3 >5.0 – 50 spins)
1	Lop-sided, All <5 or >5 (10), All <21 or >30 (50)	2, 1, 3, 0, 4, 0 (10 spins) 1, 3, 7, 8, 9, 11 (50 spins)
0	Inappropriate response	25, 50, 75, 25, 50, 75 (out of range)

Table 4: Codes and examples for Q2 to Q5 completed by grades 3, 5, 7, and 9

Suggestion of appropriately surprising outcomes in Q4 increased from 42.7% of grade 3 students to 76.6% of grade 9 students. There was, however, a monotonic decline over the grades in the ability to provide reasonable variation in the six suggested outcomes of trials in Q5, with 36.0% in grade 3, 32.6% in grade 5, 22.3% in grade 7 and 18.2% in grade 9 doing so. In fact across the grades 1.3% of grade 3, 15.5% of grade 5, 14.1% of grade 7, and 17.2% of grade 9 suggested outcomes in strict accordance with theoretical probability, e.g., 5, 5, 5, 5, 5, 5 (10 spins) or 25, 25, 25, 25, 25, 25 (50 spins). At the same time the percent predicting too much variation fluctuated with 20.7% in grade 3, 14.4% in grade 5, 31.5% in grade 7, and 26.0% in grade 9 doing so.

Table 5 contains the means and standard errors for all grades for Sample 1 on the PEVar subscale. There was a steady rise in performance from grades 3 to 7, with a plateau evident between grades 7 and 9. There were significant differences between grades 3 and 5 ($p < .001$) and between grades 5 and 7 ($p < .01$) on the PEVar subscale.

	PEVar				DisVar	
	G3	G5	G7	G9	G7	G9
Mean	4.15	5.28	5.94	5.81	6.18	6.29
Std Error	0.203	0.172	0.154	0.153	0.203	0.201

Table 5: Means and standard errors for Sample 1 on the variation subscales

As can be seen in Table 6 all grades in Sample 2 showed a significant improvement on the PEVar subscale from the pre- to the post-test ($p < .001$ – grades 3, 5, and 7; $p < .01$ – grade 9). Again a plateau effect is evident from grades 7 to 9 on the PEVar subscale for the post-test after a rise in performance from grades 3 to 7 on these items.

		PEVar				DisVar	
		G3	G5	G7	G9	G7	G9
Pre	Mean	3.89	4.93	5.81	6.12	5.93	6.30
	Std Error	0.336	0.269	0.224	0.217	0.296	0.280
Post	Mean	5.19	6.24	6.97	6.79	8.80	7.20
	Std Error	0.282	0.192	0.226	0.191	0.290	0.271

Table 6: Mean and standard errors for Sample 2 on the variation subscales

Code	Summary	Examples
Q10. How would you describe the shape of the graph?		
3	Acknowledges variation	Most of it land landed on 20; Jaggedy; Up and down
2	Reasonable shape description	Pyramid; City; Hill; Melbourne (Physical objects) Triangle; Rectangular; Circle (Geometry)
1	Focuses on graph and axes	Line graph; Column graph (Graph types) Straight; Flat; Even; A line (Axes)
0	Unreasonable responses	Small; Strange; Different; Big (Illogical); Don't know (NR)
<i>Q11-Q13. Imagine that three other classes produced graphs for the spinner. In some cases, the results were just made up without actually doing the experiment.</i>		
3	Right choice with appropriate reason	Q11: Made up – Shape of a triangle Q12: Made up – The range is too big Q13: Real – Not even but around 25
2	Right choice with vague reason	Q11: Made up – It would be impossible Q12: Made up – Doesn't look real Q13: Real – More stable graph
1	Right choice with no reason Wrong choice with data based reason	Q11: Made up – ? / Real – Around 25, the average Q12: Made up – They are! / Real – All over the place Q13: Real – Don't know / Made up – All in one spot
0	Wrong choice with no reason, illogical or vague reason	Q11: Real – They would not lie Q12: Real – It's random Q13: Made up – It looks wrong

Table 7: Codes and examples for Q10 to Q13 completed by grades 7 and 9

Distributional Variation

Table 7 contains examples of responses to the DisVar subscale answered by students in grades 7 and 9. Neither grade showed very much imagination (code 1 or above), in answering Q10, with 13% of grade 7 and 16.7% of grade 9 mentioning something to do with variation. For the three questions asking about real or made up distributions, 73.1% of Sample 1 students judged appropriately with either vague or appropriate reasoning, for Q11 (codes 2 or 3). For Q12, this percent dropped to 40.2% indicating more difficulty in

appreciating too much rather than too little variation. For the reasonable distribution in Q13, 64.1% of students gave vague or appropriate reasons in supporting this view. For Sample 2, the percents before and after instruction for Q11 were 73.9% and 86.7%, for Q12 were 36.1% and 57.2%, and for Q13 were 63.3% and 77.2%. This indicates that after instruction students were more easily able to identify appropriate and inappropriate distributions by giving data-based reasons in their arguments.

Table 5 for Sample 1 shows that there is no significant difference between grades 7 and 9 on the DisVar subscale. Table 6 reveals that for Sample 2 both grades improved significantly after instruction (G7, $p < .001$; G9, $p < .01$). Although grade 9 students improved significantly on the DisVar subscale, the grade 7 students not only improved significantly, but also performed significantly better than the grade 9 students ($p < .001$) on this subscale after instruction.

DISCUSSION

Three aspects will be covered in relation to the implications of the results: outcomes of the study, limitations of the study, and educational issues for the understanding statistical variation.

The three subscales based on Figure 1 allowed the presentation of outcomes to answer the research questions in relation to prerequisite knowledge, point estimate variation, and distributional variation. All three aspects will be important in future research and classroom planning. Teachers need to be aware, for example, of initial unfamiliarity with “chance” by young children and the different degrees of recognition of lowest and highest values in a graph by middle school students; of the lack of spontaneous acknowledgement of variation without prompting; and of the apparent greater difficulty of recognizing too much variation as inappropriate compared to too little variation. Tasks such as Q5 and Q11 to Q13 should provide useful diagnostic information for teachers about students’ beliefs regarding the extent of reasonable variation. The instances where students predict no variation in six repeats of many trials in Q5 (e.g., 12.6% overall in Sample 1) or claim that the distribution in Q11 is likely to be real rather than made up (e.g., 13.6% overall in Sample 1) suggest a lesson in the teaching of probability: Expectation must be balanced by variation. On the other hand, the instances of providing too much variation in six repeats of many trials in Q5 (e.g., 23.3% overall) or not realizing that the distribution in Q12 was made up rather than real (e.g., 22.3% overall) point to another issue for the classroom: Students need many hands-on experiences to develop an appreciation for “how much” variation is reasonable. This will not happen in a single lesson.

Limitations of the study from a measurement perspective are associated with using 10 repeated spins for grades 3 and 5 in Q2 to Q5, and 50 spins for grades 7 and 9. This may account for the fluctuating performance across grades for Q5. Having little control over the teaching arrangements in grades 7 and 9 may have contributed to greater improvement in performance on the PEVar and DisVar subscales for grade 7 than grade 9. The plateau of performance at grade 9 may also reflect the classes chosen for participation by the schools or an overall lack of continuing interest in chance and data by schools over the middle years.

Variation is everywhere and has different manifestations depending on where it occurs and what causes it. The choice of scenario found in this paper was intended to simplify some aspects of variation in order to concentrate on the statistical aspects of variation from a point estimate and variation from the theoretical distribution of random outcomes related to the point estimate. Children as young as six have little trouble appreciating that variation occurs in chance settings (Kelly & Watson, 2002), however, what is not easy is finding tasks that can determine to what degree students imagine variation occurring. Asking about repeated trials is a starting point. It would appear, however, that some type of visual presentation, probably graphical, is required before the more complex issue of appreciation of variation from distributions can be explored. It will be interesting to follow future research as other tasks are developed to investigate student understanding.

Acknowledgement. Funded from Australian Research Council grant no. A00000716.

References

- Kelly, B. A., & Watson, J. M. (2002). Variation in a chance sampling setting: The lollies task. In B. Barton, K. C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds), *Proceedings of the 25th annual conference of the Mathematics Research Group of Australasia: Mathematics education in the South Pacific, Auckland: Vol. 2* (pp. 366-373). Sydney: MERGA.
- Konold, C., & Higgins, T. L. (2002). Working with data: Highlights related to research. In S. J. Russell, D. Schifter, & V. Bastable (Eds.), *Developing mathematical ideas: Collecting, representing, and analyzing data* (pp. 165-201). Parsippany, NJ: Dale Seymour Publications.
- 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. Pre-session Research Symposium conducted at the 77th Annual National Council of Teachers of Mathematics Conference, San Francisco, CA.
- Torok, R. (2000). Putting the variation into chance and data. *Australian Mathematics Teacher*, 56(2), 25-31.
- Watson, J. M., & Kelly, B. A. (2002a). Can grade 3 students learn about variation? In B. Phillips (Ed.), *CD of the Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society, Cape Town, South Africa*. 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), *Proceedings of the 25th annual conference of the Mathematics Research Group of Australasia: Mathematics education in the South Pacific, Auckland: Vol. 2* (pp. 682-689). Sydney: MERGA.
- Watson, J. M., & Kelly, B. A. (2002c, December). *School students' understanding of stacked dot (line) plots*. Refereed paper presented at the Australian Association for Research in Education conference, Brisbane.
- Zawojewski, J. S., & Shaughnessy, J. M. (2000). Data and chance. In E. A. Silver & P. A. Kenney (Eds.), *Results from the Seventh Mathematics Assessment of the National Assessment of Educational Progress* (pp. 235-268). Reston, VA: NCTM