Predicting Dice Outcomes: The Dilemma of Expectation Versus Variation

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

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

This study considers students' predictions and explanations for outcomes when a normal six-sided die is tossed 60 times. Changes are monitored after lessons on chance and data and/or after two years. The study is motivated by two approaches to probability advocated in the mathematics curriculum: one stressing expectation based on theory and the other acknowledging variation during experiments. The outcomes are discussed in light of other research and the dilemmas created for students by these two approaches to probability.

A National Statement on Mathematics for Australian Schools (Australian Education Council [AEC], 1991) is quite explicit in its objectives related to chance. In Band C, for example, students are to be given experiences that will enable them to

> C2: construct sample spaces to analyse and explain possible outcomes of simple experiments and calculate probabilities by analysis of equally likely events
> C3: estimate probabilities using the long-run relative frequency (that is, empirical probabilities). (pp. 175-176)

The dilemma explored in this report is stated under the second point above in the following possible activity for students.

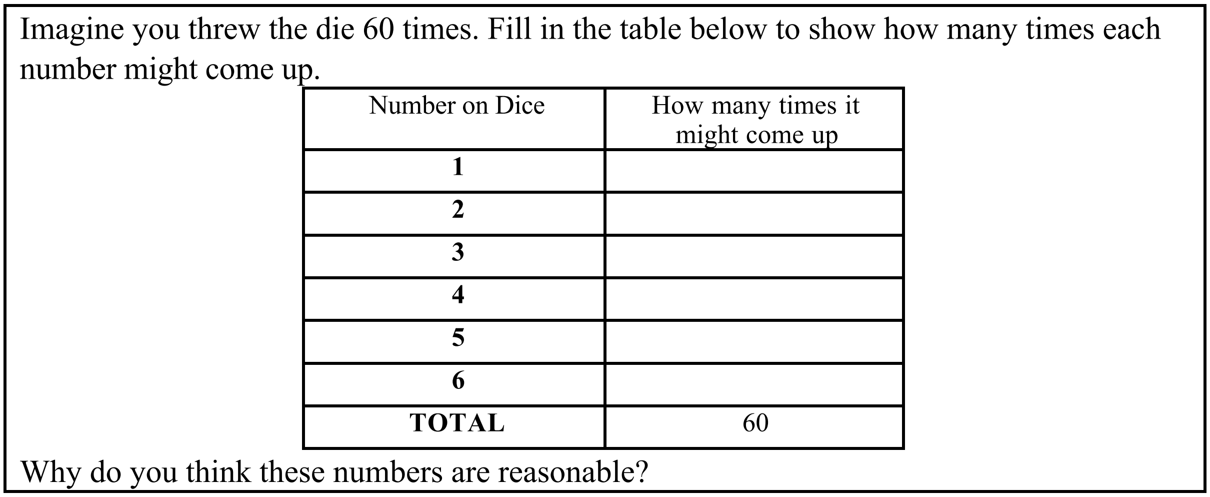
> Recognise that while the analytic probability of, for example, tossing a 6 on a die may be one-sixth, this does not mean that one 6 will appear in every six tosses but rather that in the very long run 6 will appear roughly one-sixth of the time. (p. 176)

What is acknowledged, although not explicitly stated in the above activity, is the presence of variation in the random process associated with tossing a die.

The increased focus on variation as the central issue in statistical thinking and the teaching of statistics (Moore, 1990; Wild & Pfannkuch, 1999) places great demands on students in balancing variation with expectation when dealing with probability. The theoretical probability predicts the most likely outcome or a series of equally likely outcomes, as in the case of fair dice, but cannot guarantee these outcomes in repeated experimental trials. Suggested activities for students tend to focus on the decreasing variation from the expected outcome as the number of trials increase, but there is little recognition of the need to discuss appropriate variation in a moderate number of trials. In an early study exploring this aspect of expectation versus variation, Green (1983) presented students with three bar graphs representing the outcome of 60 tosses of a fair die. One graph had a peak in the middle, one graph was an exact uniform distribution, and one graph presented frequencies varying between 8 and 12 for the six outcomes. Although the largest group of students chose the graph with appropriate variation to represent the 60 tosses, Green was concerned that 36% of students picked the peaked graph and 20% chose the theoretical even distribution.

In findings of more recent studies of students' strategies for handling expectation and variation, there is a tendency for students to predict individual outcomes based on theoretical proportions, but to include some degree of variation in predictions of repeated trials (Kelly & Watson, 2002). In predicting how many red objects in a handful of 10 drawn from a container of 100 objects, 50% of which are red, for example, students are likely to say “5”, but when predicting six repeated trials (with replacement), are more likely to say “5,8,4,5,7,6”. The exception to this is seen for a few secondary students who predict the most likely outcome for all repeated trials based on their study of probability, indicating that among all possible outcomes this is the most likely (Reading & Shaughnessy, 2000). These studies, however, consider predictions of simple independent outcomes in terms of a single theoretical proportion, whereas Green’s scenario and the one employed in this study consider the prediction of a distribution of all possible outcomes for a sample space, with these being dependent on each other. How the context may affect outcomes is of interest if similar tasks are to be developed for classroom use.

The task used as a basis for this study is shown in Figure 1 and is a variation on the one used by Green (1983), asking students to generate their own distributions rather than identify the most appropriate one. This was the second question used on a survey of students’ understanding of variation in relation to the chance and data curriculum (Watson, Kelly, Callingham, & Shaughnessy, 2003). The first question on the survey was a lead-in to the task in Figure 1, asking students which outcome was more likely when a normal 6-sided die was tossed: a one, a six, or the two were equally likely. Justification of the response was also requested.



Imagine you threw the die 60 times. Fill in the table below to show how many times each number might come up.

Number on Dice	How many times it might come up
1	
2	
3	
4	
5	
6	
TOTAL	60

Why do you think these numbers are reasonable?

Figure 1. “60 tosses of a die” survey item

Of interest in analysing the responses to this task were (1) students’ basic understanding of the task in accounting for 60 tosses of the die, (2) students’ appreciation of the variation from a strict probabilistic interpretation of outcomes due to the random nature of the process, expressed in the numbers given, and (3) students’ explanations of their numerical values in terms of strict probability or variation. Due to the use of the task in Figure 1 as part of a larger research study it was also possible to consider differences across grades 3 to 9, students’ change in performance after some classroom chance and data experiences that were devised to enhance appreciation of variation, and their change in perceptions after a period of two years.

Methodology

Sample. For ease of description, the sample of students who participated in this study are described as Sample 1, Sample 2A, Sample 2B, and Sample 3. Students in Sample 1 were from 10 government schools in Tasmania in grades 3, 5, 7, and 9. Sample 2A was a subset of the students in Sample 1 who completed lessons on chance and data with an emphasis on statistical variation and who also answered the task in Figure 1 as part of a post-test. Students in Sample 2A in grades 3 and 5 experienced ten lessons taught by an experienced teacher as discussed in Watson and Kelly (2002a) and students in Sample 2A in grades 7 and 9 had a unit of work taught by their usual mathematics teachers as described in Watson and Kelly (2002c). None of the lessons, however, employed the task presented in Figure 1. Sample 2B, a further subset, consisted of the students who were again re-tested with the same survey task two years later. Sample 3 was a disjoint subset of Sample 1 who answered the same survey task two years after the original testing took place but did not complete specialised lessons on chance and data as part of the research project. Students in Sample 2B and 3 were in grades 5, 7, 9, and 11 when they were re-tested. Table 1 contains the numbers of students in each sample for each of the grades.

Table 1
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	93	91		338
Sample 2B		47	58	66	28	199
Sample 3		67	44	68	31	210

Analysis. The criterion for determining the appropriateness of variation displayed in the numerical responses was based on a simulation of 1000 outcomes for 60 tosses of a die using an EXCEL spreadsheet. The standard deviation for each simulation was calculated and values were ordered. Appropriate variation was determined by the standard deviations falling within the middle 90% of the results. This meant that appropriate variation was demonstrated if a response's standard deviation fell between 1.2 and 4.7.

The hierarchy employed in developing the coding scheme reflected the appropriateness of the various components of the response in terms of showing an understanding of chance variation and the structure of the overall response in terms of combining components in a meaningful fashion. The latter structure reflected aspects of the SOLO Taxonomy (Biggs & Collis, 1982; Biggs, 1992) in that successively more elements of the problem (set in the concrete symbolic mode) were combined in explanations: Prestructural responses had difficulty understanding the task, Unistructural responses reflected one element of the task sometimes not acknowledging conflict, Multistructural responses included more than one element but did not fully resolve the issue of variation, and Relational responses appreciated the nature of variation within the random process. The combination of statistical appropriateness with structural criteria resulted in a further level of response, labelled Transitional, between Unistructural and Multistructural. Explanations of the five codes used, with examples, are given in the Results section.

The responses from Sample 1 were the basis for the initial discussion of the results, in particular for the distribution of levels of response across grades. Paired *t*-tests were used to indicate whether improved performances had occurred for Sample 2A after a series of lessons on chance and data, for Sample 2B after a further two years of schooling, and for Sample 3 after two years but with no instructional intervention.

Results

Based on the students in Sample 1, Table 2 contains a brief summary and examples of responses to the task in Figure 1, including the percent of students who responded at each coding level. The following paragraphs describe the levels of response and trends over the grades, which are reported as percents at each level in Table 3.

Table 2
Levels of Response to the "60 tosses of a die" Task

Code	Structure	Description	Examples
4	Relational 4.47%	Appropriate variation and explanation	"12, 9, 11, 10, 10, 8 - Because they're all around the same but you can't know if they will come up that number of times."
3	Multistructural 13.96%	Conflict of probability and variation, variation with no explanation, or explanation but too much variation	"10, 10, 10, 10, 10, 10 - In theory all numbers should come up equally. They probably will not." (Realised Conflict of probability and variation) "9, 12, 10, 7, 6, 16 - I used these numbers based on what usually happens to me." (No explanation) "15, 8, 10, 2, 19, 6 - Because there is one of each so it could be any number." (Too much variation)
2	Transitional 22.76%	Strict probability or too little variation	"10, 10, 10, 10, 10, 10 - They all have the same chance of coming up." "10, 10, 9, 11, 10, 10 - These numbers are reasonable because there is a chance in six."
1	Unistructural 28.05%	Add to 60 without appropriate variation and explanation or do not add to 60 with aspect of variation	"10, 10, 10, 10, 10, 10 - It was a guess" "10, 20, 10, 5, 5, 10 - Because it adds to 60" "19, 18, 5, 7, 23, 10 - Because any number can come up."
0	Prestructural 30.76%	Do not add to 60 or have unrealistic value	"6, 3, 2, 1, 4, 5 - Because the one might have a bigger chance of coming up more than the other numbers."

Prestructural. Students who gave Prestructural responses (code 0) did not appreciate the need to account for 60 tosses of the die and hence could not demonstrate appropriate variation. As well as this, they did not display reasoning associated with variation in the task, providing none at all, idiosyncratic explanations, or reasons vaguely related to equal chances. In Sample 1, 50% of grade 3 students and 30% of grade 5 students responded in this fashion. The percent of responses with code 0 decreased with increasing grade.

Unistructural. There were two types of Unistructural responses (code 1). One type produced numbers that summed to 60 but did not show appropriate variation and had an explanation reflecting neither probability nor variation. The other type produced numbers that did *not* sum to 60 but an explanation based on an intuitive view of variation expressed in a phrase similar to "anything can happen". Around 30% of students in each grade in

Sample 1 responded at the Unistructural level, suggesting an inability to cope with more than one aspect of the task.

Transitional. Those in the Transitional phase (code 2) were those who gave strict probabilistic estimates or explanations, involving numbers that, although summing to 60, showed no variation or too little variation. Explanations were based on equal likelihood of outcomes (often expressed as "same chance") or occasionally on the geometry of the die. These responses were considered transitional based on the two general criteria used in coding noted in the Analysis section: statistical appropriateness in terms of the purpose of the task and structural complexity. Although students focused on two aspects of the task, i.e., adding to 60 and providing a theoretically appropriate reason, there was no recognition of variation, the statistical goal of the exercise. The percent of students responding at the Transitional level in Sample 1 increased with increasing grade. Few students in grade 3 responded in a Transitional manner, whereas nearly 40% of grade 9 students did.

Multistructural. At the Multistructural level (code 3), responses, all of which added to 60, included some appropriate aspect of variation, either in the appropriate estimates for the 60 trials, or in the explanation, but not both. Some responses for the 60 trials showed appropriate variation according to the criteria set but could not explain why these were reasonable, perhaps relying on personal experiences or giving no reason. Responses that gave strict probabilistic estimates or displayed too little variation nevertheless indicated that there were many possibilities. A few responses were multiples of 5 and a few displayed too much variation, while expressing an "anything can happen" reason, supporting variation in outcomes. Responses appeared to appreciate the conflict between the underlying probability governing the die and the random nature of the process but were unable to resolve it in numbers presented, or explain it adequately. The percent of students in Sample 1 responding at this level across grades was relatively stable from grade 3 to 9, reflecting a combination of an understanding of intuitive variability without the ability to apply it appropriately to both aspects of the task.

Relational. The Relational level responses (code 4) satisfied the requirement for the 60 trials with numbers displaying appropriate variation according to the criterion set and also provided an explanation reflecting the random nature of the process. In Sample 1, no student in grade 3 responded at the Relational level, however, 7% of students across grades 5 and 7 responded appropriately. A decrease was evident in grade 9.

Table 3 shows the percent of students responding at each code for each of the grades. There is a trend for older students to respond in a more theoretically correct manner (code 2), however, older students did not necessarily acknowledge variation more often than the younger students did (code 3 or 4).

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

Code	0	1	2	3	4
Grade 3	50.0	33.5	4.0	12.5	0.0
Grade 5	30.0	27.3	19.7	15.9	7.1
Grade 7	28.0	23.7	26.3	14.5	7.5
Grade 9	16.6	28.0	39.4	12.9	3.1

Using the codes from Table 2, Table 4 shows the mean code and standard error for each grade for the task. There is a significant difference between grades 3 and 5 (-5.28, $p<0.01$) and the other grades, however, no significant difference between grades 5 and 7 or 7 and 9.

Table 4

Means and Standard Errors for Sample 1

	Grade 3	Grade 5	Grade 7	Grade 9
Mean	0.79	1.43	1.50	1.58
Std. Error	0.075	0.093	0.091	0.073

For the students in Sample 2A, paired *t*-tests were conducted for each grade to note any change in performance after instruction. Table 5 gives the means and standard errors for each grade in both the pre- and post-tests. Statistically significant improvement occurred for grades 3, 7, and 9, the greatest improvement for grade 7. Overall the number of code 4 responses doubled from pre- to post-test but was still only 10% of the total.

Table 5

Pre and Post Means and Standard Errors for Sample 2A

	Grade 3	Grade 5	Grade 7	Grade 9
Pre Mean, Std. Error	0.85, 0.118	1.43, 0.140	1.48, 0.128	1.63, 0.098
Post Mean, Std. Error	1.07, 0.121	1.43, 0.130	1.97, 0.136	1.96, 0.134
t, p	-1.69, <0.05	0, NS	-3.20, <0.001	-2.27, <0.05

Students in Sample 2B were re-tested again with the original survey two years later as part of a longitudinal follow-up. Table 6 shows the means for each of the grades for each of the tests: pre-test, post-test, and the longitudinal follow-up. Pre-test means are reported again because of the reduced sample size. Paired t-tests were again conducted among the three groups to gauge change in performance. This is also shown in Table 6.

Table 6

Pre, Post and Longitudinal Means and Standard Errors for Sample 2B

	Grade 3/5*	Grade 5/7*	Grade 7/9*	Grade 9/11*
Pre Mean, Std. Error	0.91, 0.154	1.40, 0.165	1.41, 0.149	1.82, 0.206
Post Mean, Std. Error	1.21, 0.164	1.41, 0.159	1.83, 0.164	1.50, 0.221
Longitudinal Mean, Std. Error	1.34, 0.147	1.60, 0.123	2.09, 0.147	1.82, 0.186
Pre-Post Change	-1.92, p<0.05	-0.10, NS	-2.47, p<0.01	1.30, NS
Pre-Longitudinal Change	-2.11, p<0.05	-0.99, NS	-4.14, p<0.001	0.00, NS
Post-Longitudinal Change	-0.68, NS	-1.08, NS	-1.88, p<0.05	-1.67, NS

*Grade in Longitudinal Follow-up

As can be seen, there is a positive and continual change in performance over the three conditions for students who were originally in grades 3 and 7. The changes in performance for the grade 3 students from the pre- to the post-test and between the pre-test and the longitudinal follow-up were significant, whereas, the improvement in grade 7 was significant between all conditions. Students originally in grade 5 did not improve in the post-test after instruction, however, there was a non-significant improvement two years later. Grade 9 students showed a non-significant dip in performance after instruction in the post-test but an improvement from the post-test to the longitudinal follow-up, back to their original pre-test score, indicating no change over the two years.

For the students in Sample 3 who were re-tested with the original survey two years later as part of a longitudinal follow-up, Table 7 gives the means and standard errors for each grade for each test, as well as the t and p values for potential change.

Table 7

Pre and Longitudinal Means and Standard Errors for Sample 3

	Grade 3/5*	Grade 5/7*	Grade 7/9*	Grade 9/11*
Pre Mean, Std. Error	0.83, 0.130	1.43, 0.173	1.57, 0.144	1.52, 0.179
Longitudinal Mean, Std. Error	1.04, 0.143	1.79, 0.180	2.07, 0.144	1.81, 0.142
t, p	-1.14, NS	-1.63, NS	-2.58, <0.001	-1.25, NS

*Grade in Longitudinal Follow-up

There was no significant improvement for students originally in grades 3, 5, or 9 after a further two years of schooling. There was, however, a significant difference in performance for students originally in grade 7. There were no significant differences of means in Tables 6 and 7 for corresponding grades in either year.

Discussion

Results for Sample 1 provide interesting insights into students' thinking about variation. It appears that chance experiences in the classroom are increasing students' appreciation of equally likely outcomes, as shown in the increasing percent of code 2 responses with grade in Table 2. A parallel increase in appreciation of variation is not occurring (codes 3 and 4), suggesting that school experiences need to show more emphasis on variation to balance that on expectation. This is particularly important at the middle school level.

Shaughnessy, Canada, and Ciancetta (2003) reported that for 84 grade 6 and 7 students surveyed with the task in Figure 1, 55% responded with "no variation" responses equivalent to code 2 in the current study and 30% responded with reasonable variation. Both values are higher than those observed in the current study. Similar to Kelly and Watson (2002), Shaughnessy et al. also found, that students were much *less* likely to suggest repeated outcomes with no variation when predicting the number of reds in a handful of 10 objects taken repeatedly (with replacement) from a large container with a fixed percent of red. This outcome further suggests that the contexts of tasks need to be considered carefully. Although it is possible to conceive of "5,5,5,5,5,5" for "reds in a handful" as the same type of prediction as "10,10,10,10,10,10" for 60 die outcomes, in fact the contexts for imagining the outcomes are different. For the first scenario the trials are independent and each prediction can take any value between 0 and 10, whereas in the second scenario there is a constraint that the six predicted values must sum to 60. Such a constraint may lead some students to suggest "60/6=10" for all outcomes as the easiest solution without any other consideration; the solution also "confirms the theory".

Although the results in terms of improved performance for Sample 2A after instruction are encouraging for grades 3, 7, and 9, the average performance only reaches the Unistructural level for grade 3, and the Transitional level for grades 7 and 9. The doubling of code 4 responses may be a realistic expectation but one would hope for better given the starting point. The fact that an activity mirroring closely the task in Figure 1 was not included in the lessons for students may have been a factor. The authors cannot explain the grade 5 lack of improvement, except in the above terms, as overall on the complete survey the grade 5 students did improve significantly (Watson & Kelly, 2002b). After two years the students in Sample 2B performed at similar levels to students in Sample 3 (see Tables 5 and 6), indicating that the long-term influence of the teaching unit on this particular task was negligible. It would appear that more explicit and repeated recognition of both variation and expectation is needed if the goal of genuine appreciation of their relationship is to be achieved. The authors agree with Shaughnessy et al. (2003) in recommending that many tasks in a variety of chance contexts, such as drawing objects from containers, spinning spinners, and tossing dice, be repeated often over the middle years.

It is important that future research not only employ different contexts but also employ a variety of task formats, perhaps in interviews, to explore in depth students' reasoning for their predictions. In interviews it may also be possible to distinguish reasoning associated with choosing "the most likely outcome" in repeated trials from reasoning associated with suggesting a distribution of outcomes. After completing the task in Figure 1, for example, students could be asked to consider several other tables of outcomes, some of which are made up, and justify their decisions on which are legitimate (personal communication, C. Konold, 25 September, 2002). Green's (1983) graphical alternatives might then be used. Research linking classroom experiences more closely, although not identically, to assessment tasks is also necessary to gain a greater appreciation of the parallel development of ideas of expectation and variation. Finally, teachers themselves may be a useful focus of research in terms of their own understanding of expectation and variation and how it influences their classroom planning.

Acknowledgments

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

References

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

Biggs, J. B. (1992). Modes of learning, forms of knowing, and ways of schooling. In A. Demetriou, M. Shayer & A. Efklides (Eds.), *Neo-Piagetian theories of cognitive development: Implications and applications for education* (pp. 31-51). London, Routledge.

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

Green, D. (1983). Shaking a six. *Mathematics in Schools, 12*(5), 29-32.

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.), *Mathematics education in the South Pacific* (Proceedings of the 26th annual conference of the Mathematics Education Research Group of Australasia, Auckland, pp. 366-373). Sydney: MERGA.

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

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., Canada, D., & Ciancetta, M. (2003). *Middle school students' thinking about variability in repeated trials: A cross-task comparison*. Paper submitted to the 27th Conference of the International Group for the Psychology of Mathematics Education, Honolulu.

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). Grade 5 students' appreciation of variation. In A. Cockburn & E. Nardi (Eds.), *Proceedings of the 26th Conference of the International Group for the Psychology of Mathematics Education: Vol. 4* (pp. 385-392). Norwich, UK: University of East Anglia.

Watson, J. M., & Kelly, B. A. (2002c). 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.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*, 223-265.