Variation in a Chance Sampling Setting: The Lollies Task

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

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

Responses of 73 students to an interview protocol based on selecting 10 lollies from a container with 50 red, 20 yellow, and 30 green are categorised with respect to centre and spread of numerical answers and to reasoning expressed in justification of the answers. Results are compared to earlier survey research and small-scale interview studies.

The motivation for the sequence of studies of which this one is a part arose from a survey item used in the National Assessment of Educational Progress (NAEP) in the United States, known as the "Gumball Task". Although it was set in the context of drawing a number of gumballs in a chance setting, the wording of the question encouraged an exact answer (the expected value), instead of a likely range (Zawojewski & Shaughnessy, 2000). The task was rewritten in three alternative forms by Shaughnessy, Watson, Moritz, and Reading (1999) in an attempt to explore more deeply students' understanding of variation and to move away from the strong emphasis on centres. In association with trials of the rewritten tasks in surveys with 324 students in the United States and Australia, Shaughnessy et al. developed a two-dimensional coding system based on a "centring scale" and a "range scale". The "centring scale" for a set of six predictions, for example, classified responses using the mean to determine whether the response was "low", "five" (mean), or "high". The "range (r) scale", informed by a simulation of 1000 trials, resulted in a classification scheme of "narrow" ($r \leq 1$), "reasonable" ($2 < r < 7$), or "wide" ($r \geq 7$). Each response was then classified based on centre and spread, with the optimum classification being a "five-reasonable" prediction.

Two small studies followed the surveys, where students were interviewed to gain more appreciation of their understanding of the "Lollies Task", as it was renamed for Australia. Torok and Watson (2000) used a similar protocol to that used here as well as several others involving variation and conducted 16 interviews with students in grades 4, 6, 8, and 10. The responses were categorised and clustered into similar groups that formed a four-tiered hierarchy demonstrating an increasing sophistication in the understanding of the proportional ideas and the variation involved in the tasks. The categories they found were those displaying (i) "weak appreciation of variation", (ii) "isolated appreciation of aspects of variation and clustering", (iii) "inconsistent appreciation of variation and clustering", and (iv) "good consistent appreciation of variation and clustering" (p. 155). Reading and Shaughnessy (2000) also interviewed 12 students using a similar protocol to this study and reported on four case studies from one student in each of grades 4, 6, 9, and 12. These students reflected many of the characteristics of the four levels observed by Torok and Watson, with the grade 12 student expressing conflict in choosing between multiple choice responses representing strict probability and sampling variation.

Following these four studies, this report focuses on two research questions. First, following the survey work of Shaughnessy et al. (1999) what are the distributions of students' responses given in an interview setting with respect to centres, spreads, repeated values in predictions, and change in predictions after experimentation? Second, following

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. 366-373). Sydney: MERGA.

the work of Torok and Watson (2000) and Reading and Shaughnessy (2000), are four levels of understanding confirmed for the overall context of the Lollies Task? Does the structural complexity fit that observed earlier?

Method

Interview protocol. The Lollies Task was the first and longest protocol of an interview focusing on aspects of variation in chance and data in a 45-minute session. The task was based on a container with 50 red, 20 yellow, and 30 green lollies in it. The first part of the protocol was a series of questions asking for predictions of how many red lollies are likely in handfuls of 10 (see Figure 1). Students were later given the opportunity to perform the experiment for themselves and go back and change any answers already given. The second part of the protocol (not shown here) was a graphing exercise that asked the students to imagine that 40 students performed this experiment and to "draw" what this might look like. Students who had trouble producing graphs or understanding the question were presented with labelled axes to help with the task. All interviews were video taped and transcribed for later analysis.

1. Suppose you have a container with 100 lollies in it. 50 are red, 20 are yellow, and 30 are green. The lollies are all mixed up in the container. You pull out 10 lollies.

a) How many reds do you expect to get?

b) Suppose you did this several times. Do you think this many would come out every time? Why do you think this?

c) How many reds would surprise you? Why do you think this?

2. Suppose six of you do this experiment.

a) What do you think is likely to occur for the numbers of red lollies that are written down?

_____, _____, _____, _____, _____, _____ Why do you think this?

3. Look at these possibilities that some students have written down for the numbers they thought likely.

(a) 5,9,7,6,8,7 (b) 3,7,5,8,5,4 (c) 5,5,5,5,5,5 (d) 2,3,4,3,4,4

(e) 7,7,7,7,7,7 (f) 3,0,9,2,8,5 (g) 10,10,10,10,10,10

Which one of these lists do you think best describes what might happen? Why do you think this?

4. Suppose that 6 students did the experiment. What do you think the numbers will most likely go from and to?

a) From _____ (lowest) to _____ (highest) number of reds. Why do you think this?

Now try it for yourself: _____, _____, _____, _____, _____, _____

Figure 1. Part of the Lollies Task interview protocol.

Sample. The sample for this study consisted of 73 students from public schools in the Australian state of Tasmania in a preparatory class (prep) ($n = 7$), grade 3 ($n = 18$), grade 5 ($n = 18$), grade 7 ($n = 15$), and grade 9 ($n = 15$). Students in grade 3 and above were chosen based on interesting or unusual responses to survey items. The prep students were considered to be bright and articulate by their teacher who had created an innovative program in mathematics throughout the school year. The students in grades 3, 5, and 7 were considered to cover a range of average to higher ability levels, whereas the grade 9 students were selected mainly from classes considered to be of average ability.

Analysis. The criteria for categorisation of the numerical responses to questions 1 to 4 in Figure 1were the same as those used by Shaughnessy et al. (1999). Although the analysis

of reasoning took place with background knowledge of the Torok and Watson (2000) study, all student explanations to the lollies task were analysed by the authors using a clustering procedure (Miles & Huberman, 1994) based on similarities in the types of reasoning shown. For this study, only the responses to the Lollies Task were considered whereas Torok and Watson had clustered responses to the complete interview, including other protocols, for example, focusing on variation in weather.

Results

First Research Question: Initial Responses

Initial analysis of the interviews focused on the choices of centres and ranges. Outcomes for responses to the parts of the protocol where students were asked for point estimates (Q1a)), lists of 6 possible outcomes (Q2a)), a choice of seven outcomes (Q3), and a range (Q4a)), are reported by grade in Tables 1 to 3. For point estimates many students said "about" or "probably" while writing a single number. Over half of these focused on 5 as an expected outcome (see Table 1). For the listing of the 6 outcomes, approximately three-quarters of the students used "five" as the centre, with slightly more students favouring a lower centre (< 4) than a higher centre (> 5). An oscillation effect was evident across grades with a peak at grade 3, a decrease in grade 5, an increase in grade 7, and another decrease in grade 9. The results for spread show an increase in the "reasonable" category for the higher grades, with a levelling out across grades 7 and 9. Over half of the students in each grade level except prep (see Table 1), combined a "five" centre and a "reasonable" spread in their choice of 6 outcomes. The tendency to predict repeated values increased to grade 7 and declined in grade 9. Only 10% of students overall changed some of their predicted values after doing 6 trials and there was no trend over grades.

Table 2 shows that for the multiple-choice question (3), half of the students chose the "five-reasonable" response (b) across grades (slightly less for grade 9). This is consistent with the results of Q2a) for grades 3 and 5, an increase in performance for the prep students, and a decrease in performance for the students in grade 7 and 9 for the "five-reasonable" combination. Very few students changed their multiple-choice responses after completing their 6 trials.

Categorisation of responses to Q4a) was determined using the classification scheme of Shaughnessy et al. (1999). The "wide" classification always included a spread of at least 7 and numbers greater or less than 5 and hence the centre did not seem relevant. The "reasonable" grouping, however, could be separated with "low", "five", and "high". Table 3 shows the results with an increase in "five-reasonable" responses across grade levels up to grade 7 and a decline for grade 9, paralleled by a decline in "wide" responses up to grade 7, then an increase at grade 9. This is inconsistent with the results for the spread of the 6 student-generated outcomes in Q2a), and also with Q3, which offered less opportunity for a wide choice to be made. For ranges with "reasonable" variation but not a centred response, grades 3 and 5 preferred "low-reasonable" estimates, whereas grades 7 and 9 preferred "high-reasonable" estimates. Again few students changed their ranges after the 6 trials.

Table 1
Percent Distribution of Q1a) and Q2a) Responses for Each Grade and Overall

	Prep ($n = 7$)	Grade 3 ($n = 18$)	Grade 5 ($n = 18$)	Grade 7 ($n = 15$)	Grade 9 ($n = 15$)	Total ($n = 73$)
Q1a) – Single outcome						
≤ 4	29	39	17	13	33	26
5	57	39	55	67	53	53
≥ 6	14	22	28	20	13	21
Q1a) – Change after trials						
	0	0	11	0	7	4
Q2a) – 6 outcomes (centre)						
Low	14	11	22	13	27	18
Five	57	83	67	80	73	74
High	29	6	11	7	0	8
Q2a) – 6 outcomes (spread)						
Narrow	0	0	11	7	7	5
Reasonable	43	61	78	87	87	74
Wide	57	39	11	7	7	21
Q2a) – Five & Reasonable						
	14	50	55	73	67	56
Q2a) – Repeated values in predictions						
	14	44	72	80	67	60
Q2a) – Changed predictions after trials						
	14	6	0	20	13	10

Table 2
Percent Distribution of Q3 Responses for Each Grade and Overall

Choice (Q3)	Prep ($n = 7$)	Grade 3 ($n = 18$)	Grade 5 ($n = 18$)	Grade 7 ($n = 15$)	Grade 9 ($n = 15$)	Total ($n = 73$)
(a) High, reasonable	0	0	0	7	13	4
(b) Five, reasonable	57	50	50	47	40	48
(d) Low, reasonable	14	28	22	20	27	23
(c) Five, narrow	14	0	11	20	7	10
(f) Five, wide	0	17	17	7	13	12
(e) (g) Other	14	6	0	0	0	3
Change after trials	0	6	11	0	0	4

Table 3
Percent Distribution of Q4a) Responses for Each Grade and Overall

Range (Q4a)	Prep ($n = 7$)	Grade 3 ($n = 18$)	Grade 5 ($n = 18$)	Grade 7 ($n = 15$)	Grade 9 ($n = 15$)	Total ($n = 73$)
Low, reasonable	0	17	17	7	0	10
Five, reasonable	14	22	44	67	47	41
High, reasonable	14	6	11	13	13	11
Wide	71	56	28	13	40	38
Change after trials	14	6	17	0	0	7

Second Research Question: Reasoning Expressed

Students were asked to give reasons for all of their numerical responses. Often students relied on similar reasoning throughout the protocol, such as "more red", "50% red", or "anything can happen", when asked for predictions. It was hence possible to combine an

assessment of reasoning with observation of consistency or inconsistency in numerical answers to describe a developmental progression based on the 73 responses.

Level 1 – Intuitive ikonic reasoning. Of the 22 students observed to respond at Level 1, 12 students produced incompatible values for their choices of 6 possible outcomes and the range of 6 outcomes. One grade 5, for example, chose 1s and 2s, then a range of 3 to 8, and appeared confused at times. Another 8 students who gave similar explanations at this level gave ranges indicating all possible outcomes (e.g., 0 to 10 or 1 to 10), so there could be no contradiction with their chosen values. Only two students, both preps, gave compatible values for their choice of 6 possible outcomes and the range of 6 outcomes.

The reasoning expressed at Level 1 was intuitive, mentioning guessing, favourite numbers, location in the container, lollies being mixed up, numbers of lollies that would fit in the hand, "you could get any", and "wouldn't get all in one go". All of the prep students were classified in Level 1, displaying only ikonic reasoning. One prep student, however, seemed to have an intuitive idea of proportionality; to Q1a) he responded "because there's 50 and 5 like ... 10." He could not clarify this answer further, however, and when asked to justify his choices for the 6 predictions (4, 6, 5, 3, 8, 2) he said he chose them because "4 + 6 = 10, 5 + 3 = 8 ... and 2 is my second best number." Graphs generated by the students tended to be imaginative (see Figure 2), drawings of the lollies, or lists of numbers.

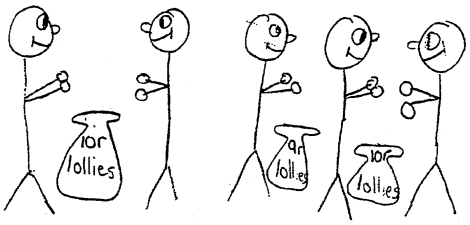

Figure 2. Level 1 representation of 40 trials.

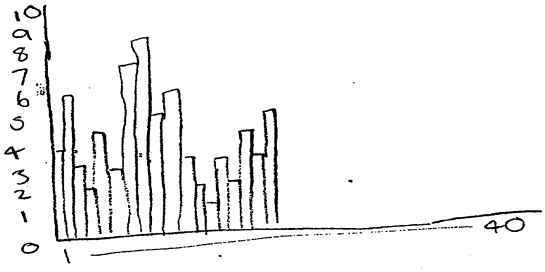

Figure 3. Level 2 representation of 40 trials.

Level 2 – "More red" but inconsistent reasoning. Of these 25 students, 7 students made low initial estimates (2 to 4) and 5 chose the Q3 response (d) that was also low (two chose (f)). Overall their six choices were low as well. Of the ranges, two were inconsistent with the six choices, one included all possible, and four were reasonable (2-6, 2-8 or 1-8). When faced with higher outcomes from their own trials, four changed some of their initial six choices to higher values. The rest of the students in this group (18) gave initial responses associated with middle or higher outcomes. Inconsistencies between the responses for the six suggested outcomes and the range occurred for two students. Five students included all possible numbers in their ranges and five others did not amend their ranges in the light of discrepant experimental outcomes.

Students at Level 2 justified their numerical responses based on "more red" without explicit mention of proportion in relation to the other two colours. One student, for example, explained his "five-reasonable" list of 6 outcomes with "because there are more reds than any other colour", and another student in response to Q3 justified her "five-reasonable" multiple-choice answer with "because there are more red than yellow and green." In the graphing task, most students generated their own representations as lollies or numbers, with a few drawing series graphs for the trials demonstrating wide variation (see Figure 3). With the help of a grid, some appeared to appreciate frequency but not centre.

Level 3 – "More" or "half" red with centred reasoning. At this level, 18 students discussed variation around the centre with non-extreme ranges consistent with their predictions. Four students in this level expressed a strong preference for the middle value and chose (c) in the multiple-choice question, with strong probabilistic reasoning, for example "half the number is red, so there is 50/50 chance of getting red." Only one student missed changing a range after the experiments were performed and four students had difficulty providing a graph consistent with their number choices and discussion.

At this level students used reasoning based on "more" or "half" red. Although, some relied on reasoning similar to that found in Level 2, an intuitive acknowledgement of "centre" was almost always present, for example, "mostly around 5 and mostly reds." Other students were more definite in their preference for "half" when explaining their numerical choices. In response to Q2a), one student reasoned that "some might go up higher, some might be lower, but half of them is red." Similarly, another student, after choosing the "five-reasonable" choice in Q3 explained "(b) – All in the middle, [other alternatives] mostly all high or low". Although many students continued to have difficulty creating their own graphs, with the help of labelled axes, ten provided idiosyncratic graphs indicating variation about the middle. One is shown in Figure 4.

Level 4 – Distributional reasoning. Eight students displayed a strong appreciation for the proportion of reds in the container as well as variation, giving reasons to Q1a) like "about 5, because there's 50 red, so divide 100 by 10 and then divide 50 by 10 to get 5." Seven could represent this in a distribution of outcomes, one of which, however, was an idiosyncratic graph by a grade 3. Another grade 3 responded appropriately to Q1a) by stating "5, because 50 is half of 100, so 5 is half of 10" and backed up his "five-reasonable" prediction of 6 outcomes by saying "they're all around the 5 mark", but could not understand the graphing task. One grade 9 chose c) in the multiple choice, because "the average number would be 5, so the most likely", but drew an appropriate distribution with explanation with variation. Most of the distributions drawn were wider than statistically appropriate, with the exception of one student who provided the tally in Figure 5. Given the lack of large-scale experimentation by the students, they showed an adequate understanding of the relationship of centre and variation in the task.

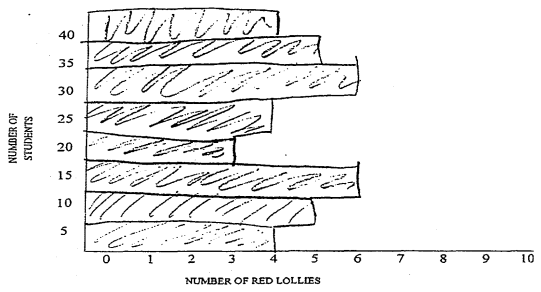

Figure 4. Level 3 representation of 40 trials.

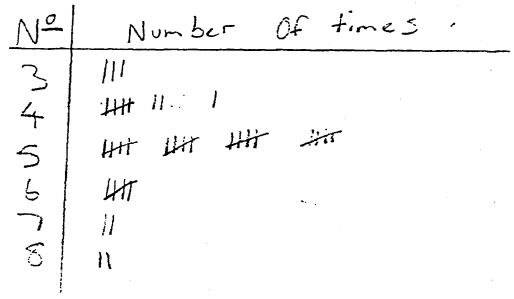

Figure 5. Level 4 representation of 40 trials.

Overall, the distinguishing features among the levels were related to appreciation of the relationship between the centre and spread of predicted values. All students understood the task and most had intuitions about both middle and spread. At Level 1 these intuitions could not be further supported with appropriate reasoning, and at Level 2 there was an attempt at justifying choices but this was inconsistent over different questions. Level 3 responses were more consistent in terms of both centre and spread but lacked strong proportional and/or distributional thinking. This was demonstrated at Level 4. Table 4

shows the grade levels of the students at each developmental level. There was a trend for increasing level of performance up to grade 7 but a drop in grade 9.

Table 4
Percent Distribution of Reasoning Levels for Each Grade and Overall

	Prep ($n = 7$)	3 ($n = 18$)	5 ($n = 18$)	7 ($n = 15$)	9 ($n = 15$)	Total ($n = 73$)
Level 1	100	44	39	0	0	30
Level 2	0	28	56	20	47	34
Level 3	0	17	6	60	33	25
Level 4	0	11	0	20	20	11

Discussion

Initial responses. Although this study used questions similar to Shaughnessy et al. (1999), outcomes were slightly different. Shaughnessy et al. had an "unclear" category for each question, for responses that were ambiguous or not appropriately answered. In the present study, a higher percentage of students were classified in the "five" and "reasonable" categories, probably a result of using an interview protocol where students read and talked through the question carefully, resulting in no missing data. The interviewer was also free to probe students' thinking to clarify nebulous comments. This was likely to contribute to a higher percent of optimal responses per grade level when compared to the Shaughnessy et al. study. The interview setting also provided the opportunity for students to express intuitive ideas of variation, for example "about 5". The fact that they only recorded the number "5" on the answer sheet probably suggests what occurred for many students in the earlier surveys. Compared to Shaughnessy et al., who found that students completing a new survey after watching the 6 trials performed in front of the class showed greatly improved responses, students in this study, performing the experiments themselves, generally were not influenced to change their predictions. For many responses, more than in the survey study, no change was required; however in other cases, change would have been appropriate. With their original work was in front of them, not starting over, perhaps a sense of "ownership" precluded some students from changing their minds.

Because this study interviewed grade 9s chosen from classes of average ability, it is not surprising that there was a decrease in "five-reasonable" responses for grade 9 in Q2a) and Q4a). For Q3 in the current study, there was a uniform performance across the grades, with a slight decrease for grade 9. The Shaughnessy et al. (1999) study, however, using combined data from all versions of the task, found fluctuations in performance across all grades, with an increase for grade 9, and a decrease for grade 12.

Both Shaughnessy et al. (1999) and Zawojewski and Shaughnessy (2000) reported larger percents of responses classified as "high" as opposed to "low" across the different versions of the task. In the current study "low" was more common than "high" in both the list and choice versions of the task. Like Shaughnessy et al. (1999), however, "wide" was preferred to "narrow" on all versions of the task. In particular the highest percents of "wide" responses in the range version of the task (Q4a)), were reported for the prep and grade 3 students, whereas for Shaughnessy et al., this was prevalent throughout all grades.

Reasoning levels. The graphing component of the Lollies Task was not given the same weight as verbal reasoning when classifying responses into the four levels described. Because it was felt that some students lacked graphing ability (especially those in prep, grades 3 and 5), their explanations of the predicted outcomes in terms of centre and spread

were weighted more heavily than technical graphing ability. For others, however, the graphing component acted as a supplement to help clarify uncertain or confused responses.

In contrast to Torok and Watson's (2000) observation of younger students being easily swayed by experiments, here most students seemed unaware that the experiments should be considered. Similar to the earlier study, students at Level 1 could not produce meaningful graphs (one pictograph that was drawn could not be explained) and there was no evidence of proportional or even majority reasoning. At Level 2 in the current study, reasoning expressed as "more reds" reflected the weak understanding of proportional ideas found earlier. Again, inconsistency was a common feature of responses about chance outcomes. For many, producing graphs was still difficult. At Level 3 students were more likely to suggest proportional ideas, such as "half", in both studies, although in the current study, responses were stronger in clustering about the middle, in both verbal suggestions and graphical representations. At Level 4, students in both studies had firm ideas of proportion and variation with clustering about the middle, but spread was often too great. Although a few students reflected the strict probability views also expressed by the grade 12 student of Reading and Shaughnessy (2000), all displayed an appreciation of sampling variation elsewhere in the protocol. Overall the increasing structural complexity with higher levels observed in earlier studies was confirmed in the current data.

This study adds to the evidence about students' understanding of sampling variation in a chance context, particularly in allowing for reasoning to be expressed, supplemented in some cases by graphs, in justifying numerical predictions. Further research is needed to consider the effect of different proportions of red lollies in the container, possible cognitive conflict if the outcomes do not match the proportions students are *told* are in the container, and the effect of collaboration if pairs or triples of students engage in the task.

Acknowledgments

This research was funded by an Australian Research Council grant (No. A00000716). The authors thank Professor Mike Shaughnessy for helpful comments.

References

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook.* (2nd ed.). Thousand Oaks, CA: Sage Publications.

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., 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*(2), 147-169.

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: National Council of Teachers of Mathematics.