MIDDLE SCHOOL STUDENTS' INFORMAL INFERENCES ABOUT FALLING RAINDROPS

Dr. Dan Canada
Department of Mathematics, Eastern Washington University, Cheney WA 99004 (USA)
dcanada@ewu.edu

The purpose of this paper is to report on the emerging results of a project that used an instructional intervention designed to improve middle school students' informal expectations of variability in a two-dimensional context that was based on the theorized placement of falling raindrops. Specifically, one aim of the project was to compare how students reasoned about variability to make informal inferences both before and after modelling a task physically and then via computer simulation. A simultaneous goal was to have students pursue their own additional questions, beyond the initial prompts given, that were prompted by an analysis of the data they had gathered.

INTRODUCTION

The underlying task in this project, based on work by others (e.g. Engel & Sedlmeier, 2005; Green, 1982; Piaget & Inhelder, 1975), posits raindrops just beginning to fall across a patio of sixteen square tiles in a 4 x 4 array: Where might the first sixteen drops land? In Engel and Sedlmeier's work, using falling snowflakes as a context, their "objective was to find out how children decide between random variation and a global uniform distribution of flakes" (2005, p. 169). Using a framework that considered the degree to which student responses reflected a perspective of randomness versus determinism, those researchers found evidence across a range of tasks and grade levels that students' ability to coordinate randomness and variability seems to deteriorate with age.

Of particular interest was the call by the researchers for instructional interventions that would leverage technology (such as computer simulations) to bolster gathering experimental data in a quest to develop "students' intuitions about chance variation" (Engel & Sedlmeier, 2005, p. 176). In fact, as detailed in the next section covering methodology, the intervention in the current project reflects the first four aspects of Engel's (2002) five-step procedure: Making initial conjectures or observations of a given phenomenon, developing a model for the purposes of simulation, gathering data, and comparing subsequent results to initial predictions. The fifth step, involving formal mathematical analysis, was beyond the purview of the middle school students.

DESIGN

The main reason "falling raindrops" was used as the context for the task instead of "falling snowflakes" is because the project initially took place with 12 students in a city Tanzania, and again at a later date with 21 students in a city in Vietnam (both places where snow was generally unfamiliar). The students had some basic skills in probability and statistics: For example, they could make simple graphs of data, and talk about distributions of data in terms of centers and ranges. Also, they could discuss likelihoods and compute probabilities for simple one-stage events.

Initially, when presented with the question of "Where might the first sixteen drops land?" as described in the previous section, students made marks on a 4 x 4 grid and also wrote down why they held that view. Whole-group discussion ensued, with student opinions ranging from a more deterministic approach (i.e. expressing that each of the sixteen tiles should contain a raindrop in the center of each tile) to more of random approach (i.e. the raindrops should look like less of a discernible pattern). The nature of the discussion had similar types of thinking as reported in similar results from other researchers (Engel & Sedlmeier, 2005; Green, 1982). Some students wondered how it was possible to make any prediction since "anything can happen" or "rain can fall anywhere", while others mused about how factors like wind might influence the results.

As we transitioned to the question of "How could we model this idea?", students were very creative. Among the ideas were finding a way to "splatter" water over a grid, or other (more viscous) liquids that were easier to record a single drop. Eventually students turned to other methods like tossing coins, blocks, and even "confetti" they made from shredded newspaper (the latter actually gave a strong impression of falling snow). Some students went up to a 2nd - floor balcony to distribute their "raindrops" (many of which missed the grid entirely), tossing things out into an alley or hall, and others

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. iase-web.org/Conference_Proceedings.php

used the height of a desk, chair, or simply standing up in a room over a grid. We allowed for all kinds of different materials and different sizes of grids (as long as they comprised sixteen squares in a 4 x 4 array), with the only requirement being that students felt their modelling technique was "as unpredictable as rain". All of the physical experimentation was photographed and videotaped for further reflection.

Once sixteen token "raindrops" had landed somewhere on the 4 x 4 grid of their choice, we did provide uniform pages of identical 4 x 4 grids on paper where they could record their results, carefully marking on the recording paper what their physical model showed. We then hung the recording papers all around the classroom: Each paper recorded one "trial" of their successful toss of sixteen "raindrops". After having at least thirty trials recorded and up around the room (all on identically-sized recording paper grids), we then entered in a period of reflection: In particular, students were asked what they noticed, and what they wondered about.

In this phase of generating new questions to pursue, the first thing many students noticed was that none of the experimental results looked like the typical "one raindrop per tile, perfectly centered in each square" which so many had suggested beforehand. In fact, soon the observation arose that most if not all of the grids up on display were without a "one raindrop per tile" result (let alone the idea of being perfectly centered). This led to the obvious connection: If there wasn't "one raindrop per tile" on a grid, then by necessity there must be some empty squares on that grid. Students began to wonder how many empty squares were among their displayed experimental results: What was the most and least number of empty squares? What was the most number of raindrops in any given square?

As students tabulated different aspects they were interested in, based on the questions about the results they raised, the notion of likelihood came up by wondering what would happen if we repeated the whole experiment on another day? The language of a "batch" of results was used to describe how many trials were on display: For example, if there were thirty grids of experimental results, we just called it a batch of thirty "trials". If, at another time, we generated a new batch of thirty results, how would students think the new batch would compare to the initial batch? As an example of a specific observation, students saw in their initial batch a grid with five empty squares, which seemed surprising to them: Would we expect to see such a grid in another batch of thirty results?

During the next part of the intervention, occurring on a different day, instead of generating more data using physical experimentation, the dynamic software "Fathom" was used (Finzer, 2000). A simulation was created in Fathom that randomly placed sixteen dots on a 4 x 4 grid, as shown in Figure 1 below.

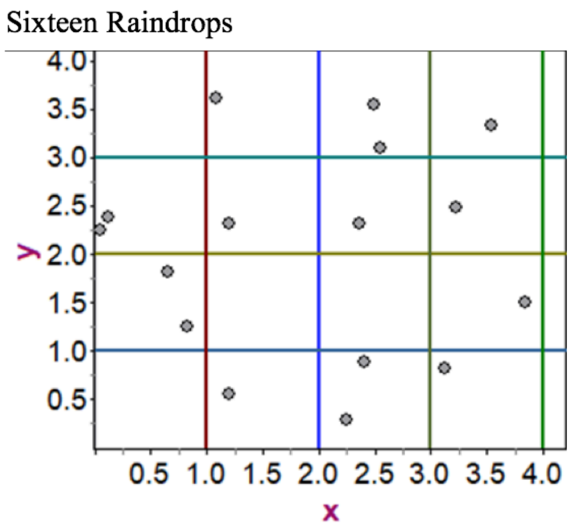

Figure 1: A single trial (via Fathom)

By toggling the animation feature, a single "trial" would unfold so the dots could slowly appear. In showing students the animation of a single trial, it was vital for students to question the veracity of the displayed result: How could they be sure the computer was doing it correctly? More salient was the question: Did the Fathom results look reasonable when compared to what the students had just done physically?

After some discussion that led to the class accepting Fathom as being just as unpredictable as their physical models, we then were able to use Fathom to look at many trials, very quickly. In fact, whereas we had previously displayed on paper a batch of at least thirty trials, we know could use Fathom to see a batch of thirty trials within seconds. The point in generating more data was to investigate some of the questions students had raised earlier, and here is where Fathom played a key role. For example, Fathom could easily record how many "raindrops" were in each tile (which were numbered 1 – 16). Figure 2 below shows the Fathom tabulation of frequencies in squares for the result corresponding to Figure 1, along with a legend showing the square labelling convention for the grid.

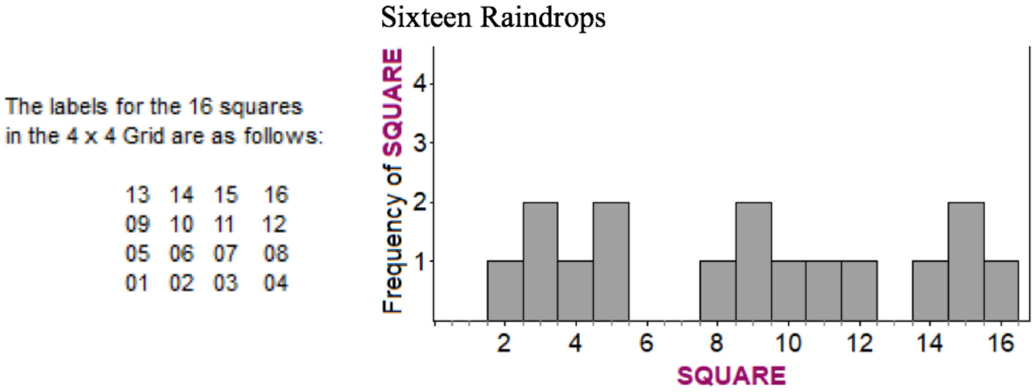

Figure 2: How many raindrops in each of 16 squares from Figure 1

At this point, the comments and questions made by students about the Fathom results paralleled those made regarding the students own experimental results. For instance, in Figure 2 we see a trial having at most 2 raindrops in any given square. How likely is such a result from any given trial? Were there any paper grids from students' own collective batch that matched that Fathom result or came close?

Again, the power of Fathom in quickly generating results came to bear as students noticed in Figure 2 that there were exactly four empty squares (they could look back at the actual grid in Figure 1 to verify that squares 1, 6, 7, and 13 were indeed empty, as cross-referenced with the labelling legend). But Fathom can record this result of "four empty squares" and then do another trial (recording how many empty squares), and so on. By using the animation feature of Fathom, we were able to slow things down in generating a batch of thirty trials, each time recording how many empty squares were in a trial.

It was important to run the initial "batch of 30 trials" on Fathom as slowly as possible, so that students could see that everything Fathom was doing mirrored the same ideas they had explored with their own paper recording grids. For example, Figure 3 shows results of such a batch of thirty trials, with frequencies for how many empty squares were in each trial.

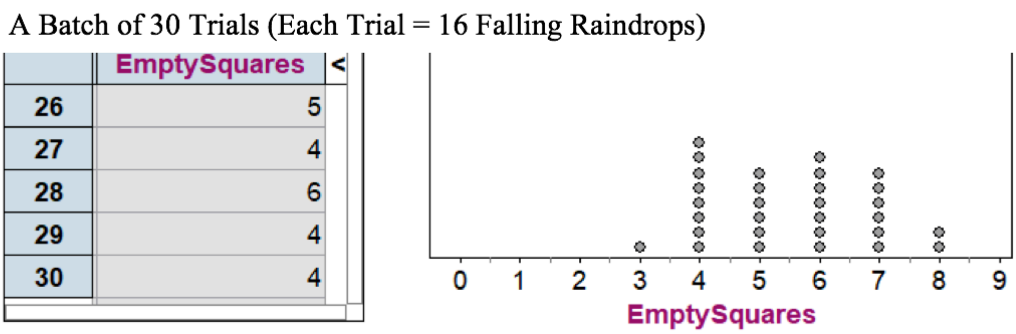

Figure 3: Counting the empty squares in each of thirty trials

The last (30th) trial had exactly four empty squares such as was seen in Figures 1 and 2. And so a tally mark (a dot in this case) was added to that column. Students could see that of the thirty trials, eight trials had happened to have exactly four empty squares. And if needed, they could go back through the other displays and match a tally mark with the grid result it came from to verify that tally mark.

By generating more data, whether in increasing the number of trials (beyond 30, for instance), or in simply replicating many batches of the same number of trials, students were able to pursue deeper questions about what was expected. They also used their insights into what was likely to make inferences about purported results. At the end of the intervention, a series of "results" of physical experimentation was given to students, which were claimed to come from a single trial, or a batch of trials, depending on what was being asked. Students were then asked to imagine that "some other class from a school across town" had submitted these "results", but we weren't sure if the other class just made up the results or if they actually came from the other class doing the physical trials. Particular attention was given to way students based their inferences of "real or fake?" on the variability inherent in the Fathom data they had just been exploring.

RESULTS

Among the questions in seeing repeated "batches of thirty trials" (which we sped up once the idea of what was going on was understood and accepted) was about what was reasonable to expect in terms of how many empty squares might be in any given trial. In Figure 3, representing a single batch of thirty trials, we see a minimum of three and a maximum of eight empty squares. So, what would be typical for the number of empty squares? If zero empty squares was considered very unlikely (corresponding to one raindrop per square), then wouldn't one or two empty squares be fairly likely?

In fact, students realized that Figure 3 was of poor use in ascertaining what was typical, since nothing too definitive emerges regarding the center of that distribution. After examining repeated batches of thirty trials, students wanted to aggregate the batches and we ended up doing 100 or more trials per batch. The time it would take Fathom to generate such results varied according to the relevant computer power, but usually something like 1000 trials only took about one minute or less. Figure 4 (below) shows the same idea of Figure 3 but a much stronger sense of distribution emerges.

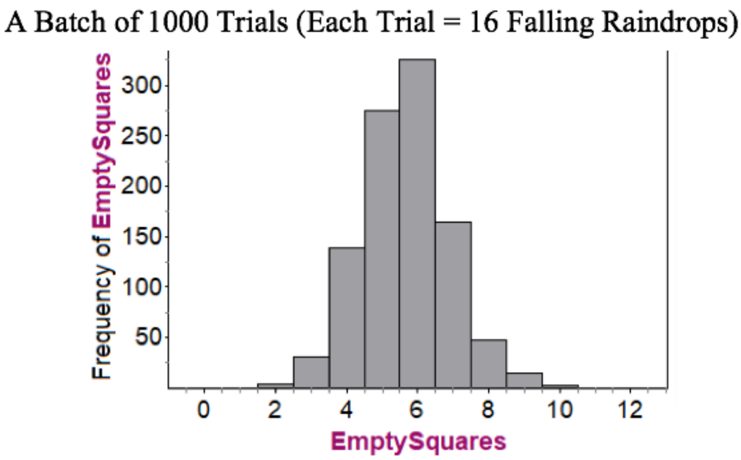

Figure 4: Counting the empty squares in each of 1000 trials

When shown some "real or fake?" data, for instance, students moved away from their claims such as "Who can tell for sure?" or "You never know, that [result] might have happened". Such claims show an over-appreciation of variability, especially when looking at the tails of distributions like that in Figure 4. Instead, as students reasoned about what was likely, they showed more sophistication in reconciling expected values with the variability they had witnessed in analyzing many results.

Especially encouraging was the way students developed new questions to help decide on what might be real or fake data, such as "How likely is it that any given trial has 6 or more raindrops on a square?" Another question that we were able to gain data on from the Fathom simulations was "How many squares in any given trial are likely to contain exactly 2 raindrops?" This manner of generating questions, along with the student comments that picked up on the variability inherent in the supporting data, was a highlight for the results of the project so far.

At this stage in the analysis of students' written responses, a rubric is being developed to better affix a quantitative measure to the degree of improvement in students' use of variability in their reasoning. For example, Engel & Sedlmeier (2005) ascribed the "Novice" label to students whose predictions included between one and three empty squares, and the label of "Expert" to those between four and eight empty squares. They then structured numeric scores based on those labels (and also the lower levels of "Deterministic" and "Moderately Deterministic"). So far, an initial qualitative examination of responses shows a general increase in student confidence in making predictions, with a markedly stronger emphasis on trying to balance variability with expected values.

DISCUSSION AND CONCLUSION

Among the surprising results of the project so far has been the new avenues for questions that came from looking at the data generated by Fathom. A good example was when students asked the wait-time question "How many trials would we expect before we hit exactly 6 empty squares?" This question seemed natural enough, given that one student after another might do a trial and not have that particular result. Or it might happen on the first try.

Some students did have bit of prior knowledge about an expected value for wait time as the reciprocal of the underlying probability, although it wasn't phrased that way. For instance, they might expect to roll a die six times to hit a "4". But again, there is variability to consider. In the context of the "falling raindrops" task, students could see that six empty squares had a high likelihood, say 0.342 for example. They then wonder if in fact $1/0.342 \approx 2.92$ might mean that "three trials" ought to be reasonable to hit exactly six empty squares. We then turned to Fathom to see if that in fact "three" was a reasonable answer for the above question on wait-time.

Perhaps the most intriguing question had to do with the probability of a square having a particular nonzero number of raindrops. They surmised a correlation between "number of empty squares" and "maximum number of raindrops in a square", but it turned out to be a challenging question to determine a specific probability for a given nonzero number of raindrops. For instance, "What's the likelihood any given trial will have 6 raindrops as a maximum on a square?" was a question that arose. Certainly we could look at our original experimental data – the paper grids up around the room – and compute that experimental probability. But getting Fathom to "keep track" of how many trials had exactly six raindrops on a square (and no more than six) was complicated for us.

Instead, it was very easy to have Fathom run trials until the number of raindrops was six or greater. So, we changed the question to "What's the likelihood any given trial will have 6 raindrops or more on a square?" To gain insight into that question, we ran 100 experiments on Fathom, where an experiment was defined as "Count how many trials are needed to be run until a trial hits 6 raindrops or more on any given square".

Before running 100 experiments, we discussed what the results might look like. Students knew an experiment could end with "one trial" because we could get 6 or more drops on a square with the first trial. Some students thought an experiment could go on for "thousands of trials" since maybe it would take a while to get the desired result. We also noted that none of our initial experimental data had that result. After discussion of initial expectations, we ran Fathom for 100 experiments as defined above, and the results are in Figure 5.

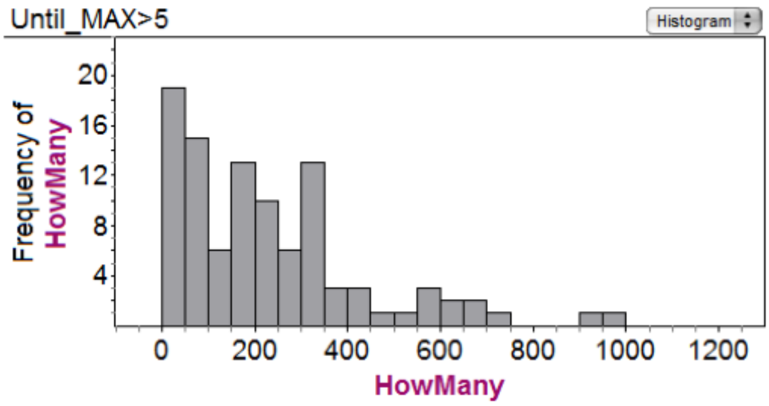

Figure 5: 100 Experiments of "How many trials to get a square with 6 or more drops?"

Using the mean result from 100 experiments, which was about 230 trials for Figure 5, students conjectured that the question of "What's the likelihood any given trial will have 6 raindrops or more on a square?" might be the reciprocal of the wait-time: $1/230 \approx 0.0043$. However, this low probability did not satisfy those who thought any given trial must surely have be fairly likely to have the desired result. Again, precise mathematical computations were not the aim of the project, but students did raise very interesting probabilistic and statistical questions. They were left musing about the correlation between "maximum number of drops" and "number of empty squares", so in that sense their curiosity had not been fully slaked.

Overall, by the end of the intervention students seemed to demonstrate three features useful for developing a more robust engagement in a world beset by variability. First, students markedly changed their predictions of where sixteen raindrops might land, as they were exposed to ever-increasing amounts of experimental data. Second, students were better able to integrate a reasoning about variability in making inferences about hypothetical results. Third, and perhaps most intriguing, students generated further questions that were based on what they noticed, and what they wondered about, in the face of large amounts of simulated data.

The latter questions are what really made this project and paper unique, in the way that students furthered their investigation of a well-known task. The next step will be to employ a conceptual framework to assess student responses in order to describe in more detail the ways in which their appreciation and use of variability improved by the end of the instructional intervention.

REFERENCES

Engel, J. (2002). Activity-based statistics, computer simulation, and formal mathematics. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*. CD-ROM.

Engel, J., & Sedlmeier, P. (2005). On Middle-School Students' Comprehension of Randomness and Chance Variability in Data. *Zentralblatt füur Didaktik der Mathematik*, *37*(3), 168–177.

Finzer, W. (2001). Fathom! (Version 1.16) [Computer Software]. Emeryville, CA: Key Curriculum Press.

Green, D. R. (1982). A Survey of Probability Concepts in 3000 Students aged 11–16 Years. In D. R. Grey (Ed.), *Proceedings of the First International Conference on Teaching Statistics*, Teaching Statistics Trust, University of Sheffield, 766–783.

Piaget J., & Inhelder, B. (1975). The Origin of the Idea of Chance in Children. London: Routledge & Kegan Paul.