THE USE OF TINKERPLOTS AND FATHOM TO PROMOTE INQUIRIES ABOUT VARIABILITY

Dan Canada
Eastern Washington University
526 5th St. (Kingston Hall 329)
Cheney, WA 99004
dcanada@ewu.edu

Matt Ciancetta
Western Oregon University
345 Monmouth Ave. N. (MNB 119)
Monmouth OR 97361
ciancetm@wou.edu

INTRODUCTION

This paper reports on an instructional intervention that involved two related tasks, both aimed at eliciting attention to, and questions about, variability in data taken from a task about *Random Walks* and a task about *Falling Raindrops.* The intervention was structured to promote middle school preservice teachers' confidence in developing and assessing criteria to make informal inferences about experimental results.

First, we give some background material for the overall intervention. Then, for each of the two tasks, the design of how the task was carried out will be described, along with the outcomes of some representative responses from the preservice teachers (PSTs) who participated. Finally, in conclusion we offer some thoughts on the implications for future classroom teaching and learning on the topic of variability.

BACKGROUND

Importance of Variability

While attempting to increase student focus on variability in data, our aim was for them to attend to the variation that arises in probabilistic situations. Within the statistics education community, variation "...does not mean an understanding of 'standard deviation' but of something more fundamental - the underlying change from expectation that occurs when measurements are made or events occur" (Watson, Kelly, Callingham, & Shaughnessy, 2003, p. 2). We live in a world filled with variation, yet much of the standard curricula at the primary grades focuses mostly on simply finding probabilities, graphing data, or finding descriptive measures such as a mean, median, or mode (Shaughnessy, 1997). Looking to collegiate math courses that prepare future teachers, developing a greater appreciation for variability can help transcend a narrow focus on descriptive statistics and bland calculations of probability.

Moreover, a report of the joint curriculum committee of the American Statistical Association (ASA) and the Mathematical Association of America (MAA) supports not only the "omnipresence of variation" as one of their core elements of statistical thinking (Moore, 1990, p. 135), but also the elements of "measuring and modeling variation" (p. 127). The "omnipresence of variability" was cited as giving rise to the very need for the discipline of statistics (Cobb & Moore, 1997, p. 801, italics in original). The idea that variability is everywhere makes sense when thinking about the world in which we live. Not only do people and their environments vary, but even repeated measurements on the same person or thing can vary (Wild & Pfannkuch, 1999). Also, "natural variation appears in the heights, reading scores, or incomes of a group of people" (Moore, 1990, p. 98).

There is also a chance variation component to our world. Moore (1990) points out that one use of probability instruction is to lead students to the understanding that chance variation, as opposed to deterministic causes, explains most outcomes in our world. Philosophically, living in a stochasticized world implies an existence beset by variation on all sides (Davis & Hersch, 1986); mathematically, "statistics provides means for dealing with data that take into account the omnipresence of variability" (Cobb & Moore, 1997, p. 801). Professional statisticians see the centrality of variation in their work, and others have framed a model of statistical thinking in which variation is the core element (Pfannkuch, 1997; Pfannkuch & Wild, 1998; Wild & Pfannkuch, 1999; Pfannkuch & Wild, 2001). The above examples lend credence to the tenet that variation is indeed the central feature behind statistics, and offer support for why others agree that "statisticians consider variation to be the foundation of statistical thinking, the very reason for the existence of their discipline" (Shaughnessy & Ciancetta, 2001).

Previous Work

To explore variability in a two-dimensional context, one idea was to use the concept of a simple *Random Walk*: While described in detail later in this paper, the basic premise was having a starting place and then taking a step either up or down with a probability of 0.50 in each case. The walk would terminate after reaching a given number of steps up or down from the initial launch. By asking participants to predict how many steps they might expect to take, and then by simulating the situation and gathering and representing experimental data, the intent was to be guided by the main elements of a conceptual framework for characterizing conceptions of variability listed below (Canada, 2006):

[1] Expecting Variation
- A] Describing What is Expected
- B] Describing Why (Reasons for Expectations)

[2] Displaying Variation
- A] Producing Graphs
- B] Comparing Graphs
- C] Making Conclusions about Graphs

[3] Interpreting Variation
- A] Defining Variation
- B] Causes of Variation

C] Effects of Variation
D] Influencing Expectations and Variation

The other idea for a task looking at variability in two dimensions was inspired by the work of others (e.g. Engel & Sedlmeier, 2005; Green, 1982; Piaget & Inhelder, 1975). *Falling Raindrops*, described in more detail later in this paper, 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, the context of falling snowflakes was used with their objecting being "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 subsequent two main sections, the intervention comprising the two tasks 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 preservice teachers (PSTs) that we were our participants in the two tasks. This paper also refers to the PSTs as students (since they were participating in a university class), if the context makes it clear.

Next, we turn to how the *Random Walks* task unfolded, and what we learned.

FIRST TASK: RANDOM WALKS

Design

Our in-class Random Walks task was based on the classic Random (Drunken) Walk scenario where an intoxicated person decides to start walking on a road or sidewalk that only runs in opposite directions, say north and south. In their drunken condition, they are equally likely to either take a step north (forward) as south (backward). Thus, there is a probability of ½ in taking a step in either of the two directions.) After each step, it is again equally likely that their next step is either forward or backward. Each of their steps are of the same length but randomly occur in either direction, north or south.

To model this phenomenon, the context of the task was switched to a two-person board game, appropriate for young students (whom the PSTs would eventually be working with in their own classrooms). This also helped create more relevance and interest among the college students. At one university in Oregon, the game was called "First to 4", and here's how the game and rules were presented to the PSTs:

This is a 2 player game with a bead on a string that is strung over a game board. The game board has a number line that goes from -4 to +4, by integers.

- *The bead starts at 0.*
- *The students take turns flipping a coin.*
- *If the coin lands Heads, then the bead is moved 1 step in the positive direction.*
- *If the coin lands on Tails the bead is moved one step in the negative direction.*
- *The game ends when the bead reaches -4 or +4 and the student with that number is the winner.*

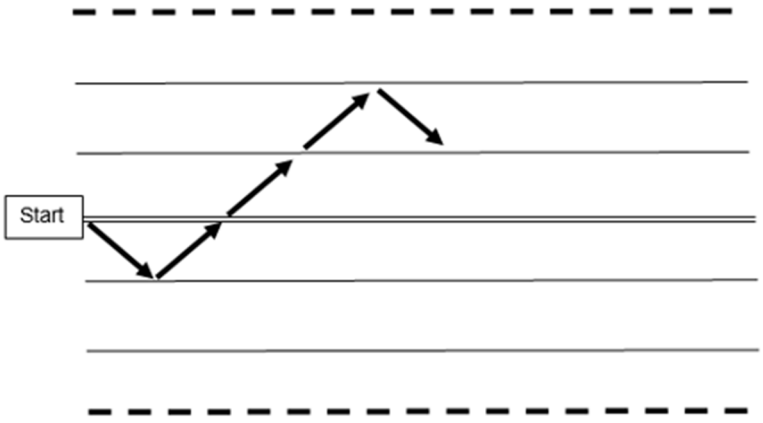
Figure 1: The "First to 4" Game

Because the expected value of the typical number of moves to terminate the game in "First to 4" might take too long to play repeated games in class, we also modified this to "First to 3" (thus shortening the typical time to play one game).

Similarly, in a university in Washington, a class of PSTs used this analogous version of a game, often called "Out of Bounds in 3" by participants. Here's how it was first introduced:

You're in the middle of a long, straight corridor, facing ahead. With you is a fair spinner that is 50% Black and 50% White. You'll take a step forward and left if the spinner lands on White, or forward and right if the spinner lands on Black.

Here's an example of what might happen in five spins:



You'll stop taking steps once you hit either of the DARK (BOLD) Dashed Lines.

Figure 2: The "Out of Bounds in 3" Game

We could easily adjust either game (the Coin-Bead version or the Spinner-Steps version) to end in 3 or 4, and while the versions were mathematically analogous, one reason for the "Out of Bounds in 3" style was because it lent itself well to the simulations we had already created using the *Tinkerplots* and *Fathom* software. By asking students ahead of time what they expected to happen, we gained an idea of their sense of variability: Someone might suggest that only four coin flips would be needed to win at "First to 4", implying no variability at all – Either repeated heads or repeated tails upon flipping the coin. Meanwhile, a suggestion of ten steps in "Out of Bounds in 3" would imply some variability in the results of repeated spins on the spinner.

The discussion on, for example, *"How many steps on average do think it might take*?", which took place before the scenario was physically acted out, was quite revealing. Some people held firmly to the idea that the steps taken in "Out of Bounds in 3" might never end, just going back and forth. Since the minimum number of steps was established as three, having the theoretical maximum of an infinite number of steps suggested that the typical number of steps was "above 3", with much disagreement on what to expect. Whereas computing an exact expected value was beyond the scope the classes we tried this in, we therefore turned to collecting experimental results.

In physically acting out the "First to 4" scenario (involving Coins-Beads), the PSTs wrote down what they noticed and what they wondered about. Much of their initial thoughts were geared toward a deterministic mindset rather than thinking in terms of randomness and variability. For example, if they "never got a negative number", then they surmised that they "weren't good at flipping". Their thoughts mirrored much of what has been written about using binary sequences of Heads and Tails, such as the use of heuristics and representativeness (Reimers, Donkin, & Le Pelley, 2018). To keep track of their results, they recorded the string of integers that the bead landed on with each move (coin flip), such as: +1, 0, -1, 0, +1, +2, +3, +4: This result was for a game that took eight total moves (coin flips) to end at +4.

Similarly, in physically acting out the "Out of Bounds in 3" scenario (involving Spinners-Steps), there was initial attention given to manipulating the spinner "fairly" so that it wasn't prone to being predictable, according to the language and thinking of the participants. Because we first assembled in small groups in the hallway, we assigned roles to people who taped off lines, who spun the spinners, who took the actual steps, and who recorded the results (on paper containing templates akin to what's shown in Figure 2). Later, after we did a few random walks in the hallway, we reconvened in the classroom to replicate some more simulated games. There, we didn't physically take the steps, but just moved a token on the recording paper so we could determine who many steps (spins) it would take to end a game.

Once the game(s) had been acted out physically, and some results had been collected, more discussion ensued. What did they notice? What surprised them? What did they wonder about? Before turning to the outcomes of these discussions in the next section, we present

an important next phase in our activities, which was to use Tinkerplots and then Fathom to run many more simulations and collect those results in an expedited way. The Tinkerplots simulation produced a result like this (for "Out of Bounds in 3"):

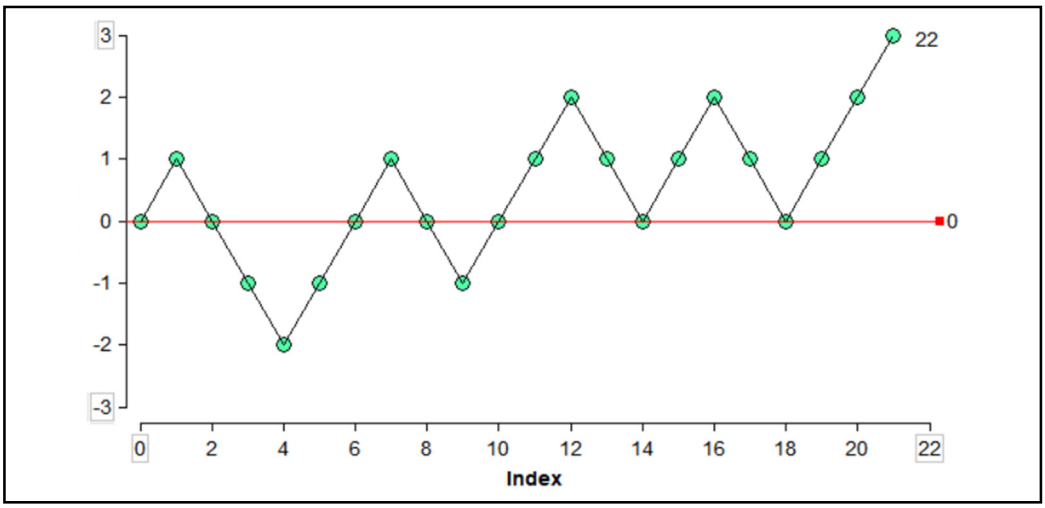


Figure 3: A *Tinkerplots* Simulation for "Out of Bounds in 3"

Note that the use of the software simulation really seemed to make sense only after having participated in the physical enactment: Even for those playing a "First in 3" game with the coins and beads, the analogy could easily be seen. In Figure 3 above, the Index along the horizontal axis indicates how many "moves" was needed. Another advantage in using *Tinkerplots* is that the graph actually spools out slowly, step-by-step, just as the participants did in the hallway. Of course, the speed of the simulation can be adjusted so that the results just appear almost instantly, but at first the students appreciated seeing the action unfold on the screen. In using the *Fathom* software, a similar style of graph is obtained (Figure 4), but there are limits to how slowly or quickly the results appear.

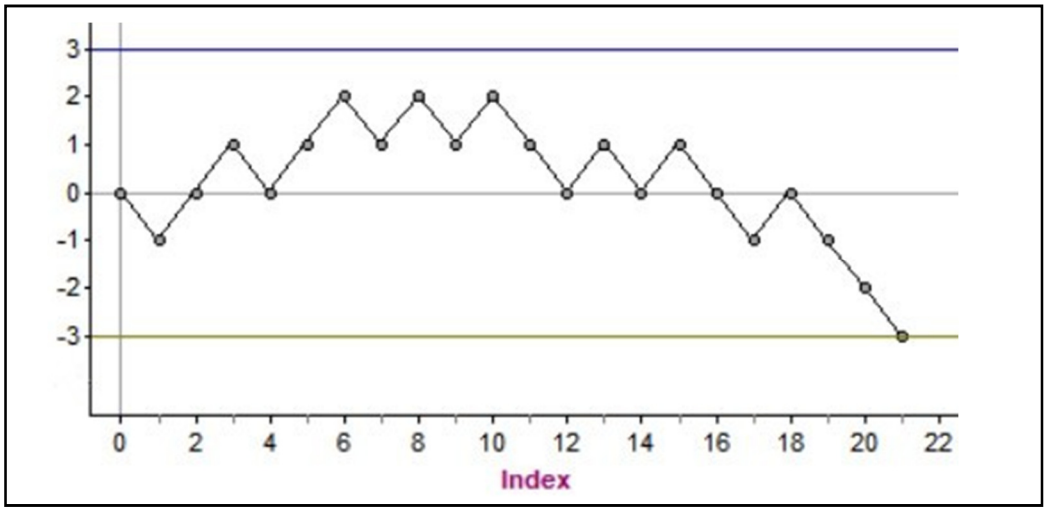


Figure 4: A *Fathom* Simulation for "Out of Bounds in 3"

Outcomes

Once students were comfortable accepting the computer simulations as being automated versions of what they themselves had performed physically, we were then able to look at far more many results than what had been collected by our classes thus far. Among the kinds of comments that emerged were how many "switchbacks" there might be, referring to the times that the direction got reversed. For example, in Figure 3 there were eight "switchbacks", which corresponded to the relative minimums and maximums along the path shown of 22 total steps. In Figure 4, there were fourteen "switchbacks", from among 21 total steps. Similarly, in what many wondered might be a correlated phenomenon, many of the PSTs wondered about how many times, on average, a path might "return to 0" – Looking again at Figure 3, we see five "returns to 0" (not counting the initial position), and in Figure 4 this was seen as six "returns to 0". Already we could see the potential in exploiting the technology to records such phenomena, and even to investigate a conjecture between how many "switchbacks" , "returns to 0", or overall number of steps a game might take. Other questions included things like "how likely is it for the bead to stay only on one side", meaning just on the positive or negative sides of the initial position.

Since our initial prompt was in terms of how many steps a game might take (or moves of the bead in that version of the game), we then collected data from many games that focused on just that parameter. Using *Fathom*, for example, we ran five quick simulations, keeping track of "how many steps" in total each game took:

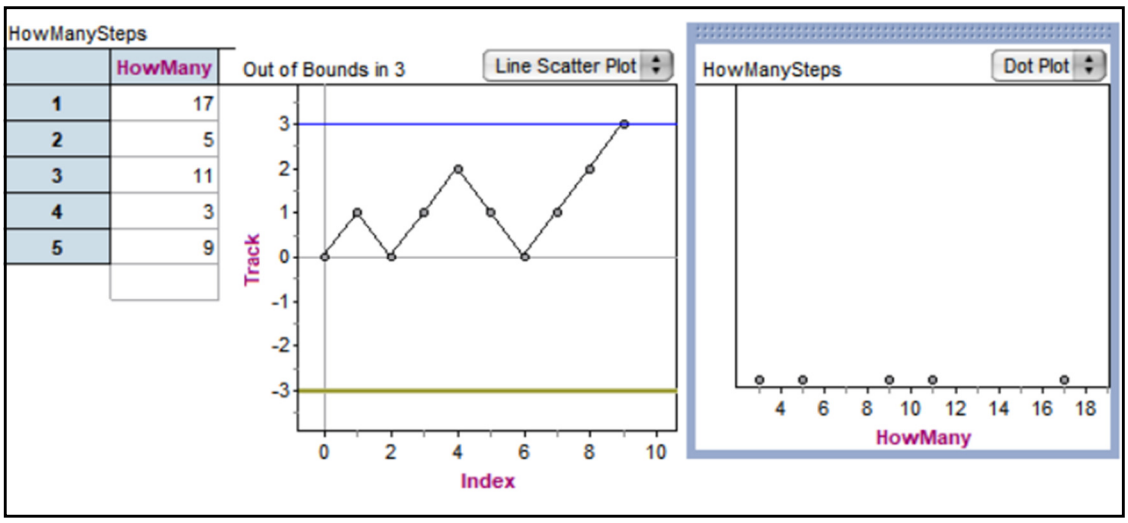


Figure 5: The results of five *Fathom* simulations

Just as we manually recorded how many total steps each of the in-class physical simulations took, so too was *Fathom* able to run multiple simulations and display the results: In Figure 5 we see that the five simulations took 17 steps, then 5, 11, 3, and the final simulation (displayed in the center) took 9 steps. Each of those results is presented in table form (displayed at the right) and in a simple dot plot (displayed at the left). Again in the spirit of inquiry about variability, student comments would include how that fifth and final simulation (displayed in the center of Figure 5) never went into the "negative

territory". After building a distribution of results from five simulations, students naturally wanted to see what many more results would look like, and here is where the power of *Fathom* really showed: It was simple to generate what we called a "batch" of 200 results:

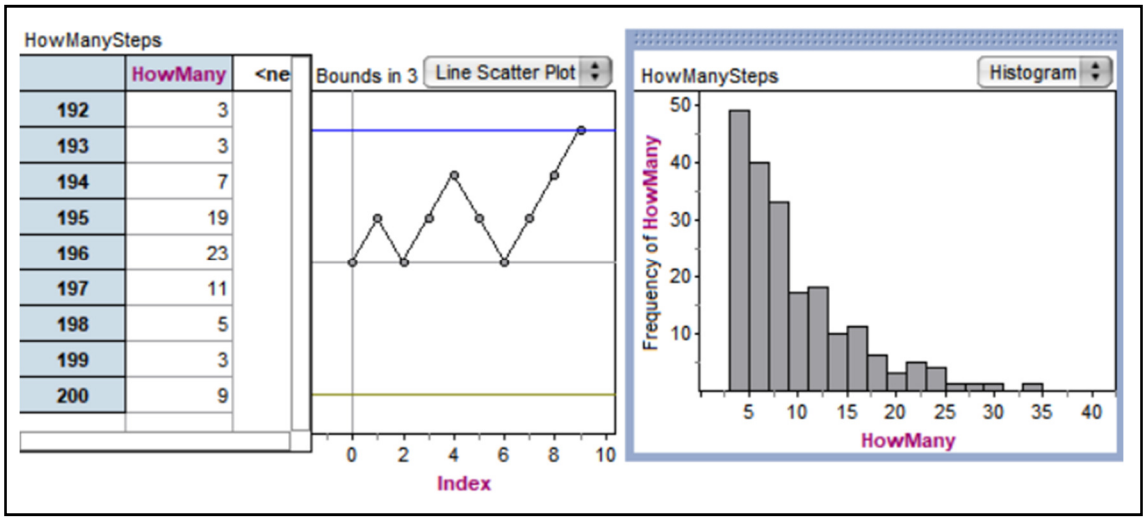


Figure 6: The results of two hundred *Fathom* simulations

Similar to Figure 5, again we see the final simulation (displayed in the center) took 9 steps. Not only is that result (of 9 total steps) the same as in Figure 5, but the actual trajectory (the step-by-step result) is identical: This is a complete coincidence, since the simulations were independently done at separate times. Nonetheless, students rightfully wondered what the likelihood is of two paths totaling 9 steps having identical trajectories. We saw this as a gateway to further explorations into probability and variation. The utility in having 200 results, however, is in giving PSTs the sense of a distribution of data. And in the case of Figure 6, we see a highly-skewed distribution: Many of the results had fewer number of steps, but what caught our students' eyes were the upper extremes.

In the end, after examining ever-larger amounts of data, students were inclined to think that "around 9 steps" would appear likely for playing "Out of Bounds in 3" (or "First to 3"), but given the variability in the results they would not be surprised at a result under 6 steps, or more than 12 steps. As a final line of inquiry, many commented on the phenomenon they called a "major switchback" : This occurred when the movement of the game (beads or steps) was within 1 of the finish mark, only to reverse course and finish on the complete opposite side of the starting position. Using "Out of Bounds in 4" as an example, Figure 7 shows what students noticed and commented upon. Although the path hits +3 at four different times, it still end up finishing at -4 . It seemed to students that, if a path was "close" in the sense of being within 1 step of ending the game, then that's the direction the path should ultimately end. As one student wrote: "Our results took many tries, & once we were close to the end, we would get pulled in the other direction." Figure 7 shows the idea of a "major switchback" (or four major switchbacks depending on how they wanted to count the multiplicities).

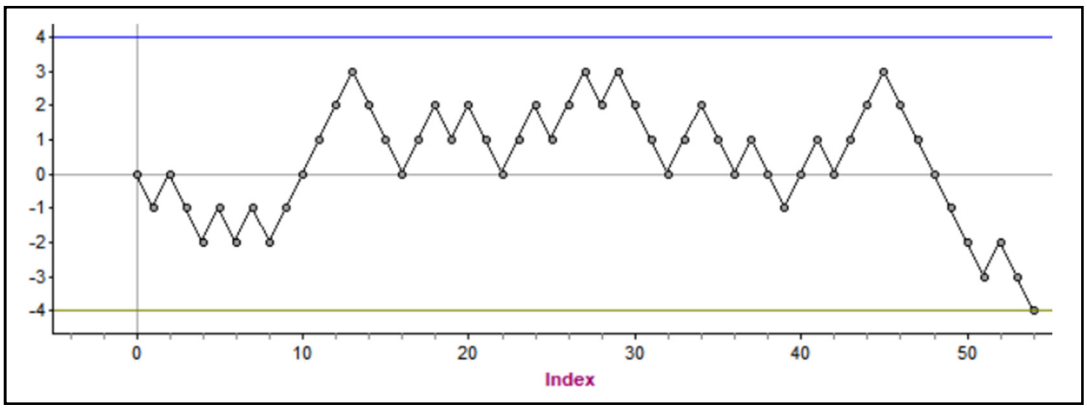


Figure 7: An example of a "Major Switchback"

Just as the use of *Tinkerplots* and *Fathom* allowed the generation of copious results from simulated data in the case of the *Random Walks* task, so too did we find great utility in exploring our second task of *Falling Raindrops.*

SECOND TASK: FALLING RAINDROPS

Design

Our in-class *Falling Raindrops* task was based on others' work profiled earlier in this paper (e.g. Engel & Sedlmeier, 2005; Green, 1982; Piaget & Inhelder, 1975): Assuming a patio of sixteen square tiles in a 4 x 4 array, where on the patio might the first sixteen drops of rain land on the tiles? It should be noted that for many iterations of doing this task around the globe (notably in Vietnam, Tanzania, Oregon, and Washington), the context of *Falling Raindrops* became "*Falling Snowflakes"* depending on weather conditions. Doing this task in the winter of the Pacific Northwest region of the United States (Eastern Washington or Western Oregon, for instance), students could be more apt to think in terms of big drifting snowflakes than of large scattered raindrops.

Initially, when presented with the question of "Where might the first sixteen drops land?", to gauge their initial expectations, 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.

To model this phenomenon using a physical simulation, 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 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".

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 *Fathom* software was used. A simulation was created in Fathom that randomly placed sixteen dots on a 4 x 4 grid, as shown in Figure 8 below:

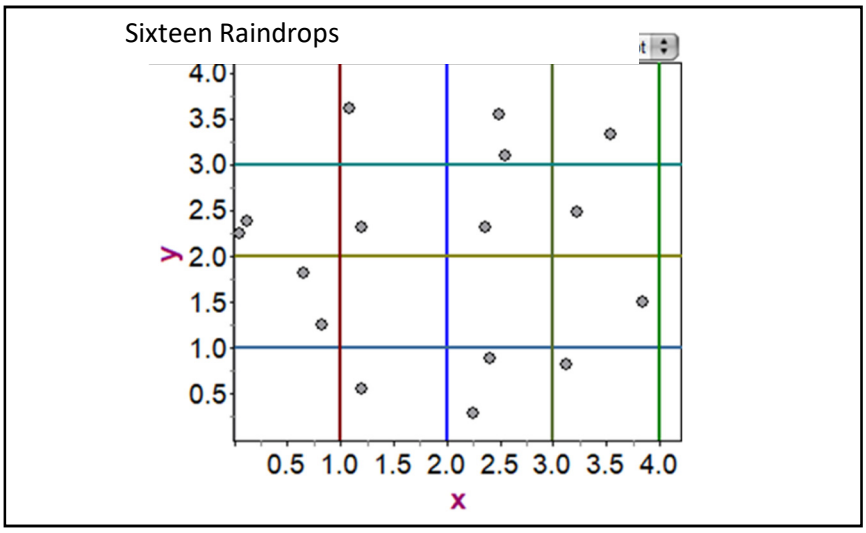


Figure 8: A *Fathom* Simulation for "Falling Raindrops"

We haven't yet figured out how to simulate this graphically on *Tinkerplots*, but only on *Fathom* so far – And while the final display of results appears quite quickly (as opposed to the gradual unfolding of results that students could emulate when doing the simulation physically), the key was 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? Connected to the many questions students had was the listing of how many raindrops were in each square, something *Fathom* could easily record (after numbering each tile 1-16). Figure 9 below shows the Fathom tabulation of frequencies in squares for the result corresponding to Figure 8, along with a legend showing the square labelling convention for the grid.

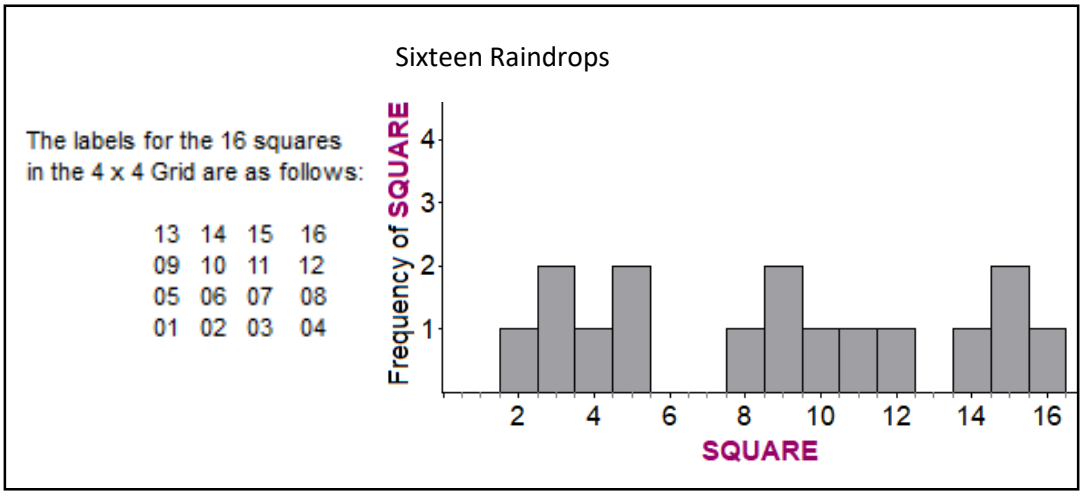


Figure 9: How many raindrops in each of 16 squares from Figure 8

In comparing Figure 8 and 9, students could see that there were indeed four empty squares, to the surprise of those who predicted a uniform distribution (i.e. one raindrop per tile or square).

Outcomes

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 (done through physical simulation), we could next use Fathom to see a batch of thirty trials within seconds. This time, for each of the thirty "trials" (each representing 16 raindrops on the grid), we had *Fathom* keep track of 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 10 shows results of such a batch of thirty trials, with frequencies for how many empty squares were in each trial.

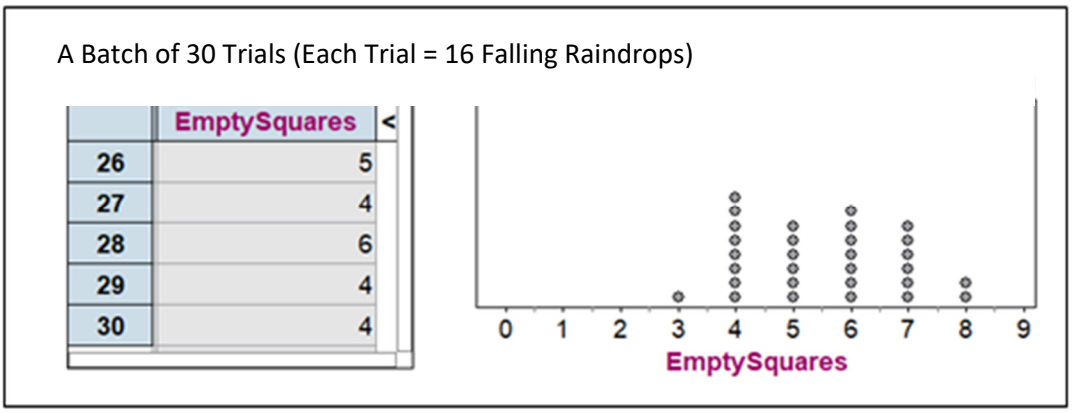


Figure 10: Counting the empty squares in each of thirty trials

In Figure 10, that last (30th) trial had exactly four empty squares such as was seen in Figures 8 and 9. And so a tally mark (a dot in this case) was added to that column, to denote four empty squares. 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. But what surprised students the most was seeing that two of the thirty simulated trials had exactly eight empty squares, and the wanted to compare that (seemingly unusual) event with the results of their own physical simulations.

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 10 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 11 (below) shows the same idea of Figure 10 but a much stronger sense of distribution emerges

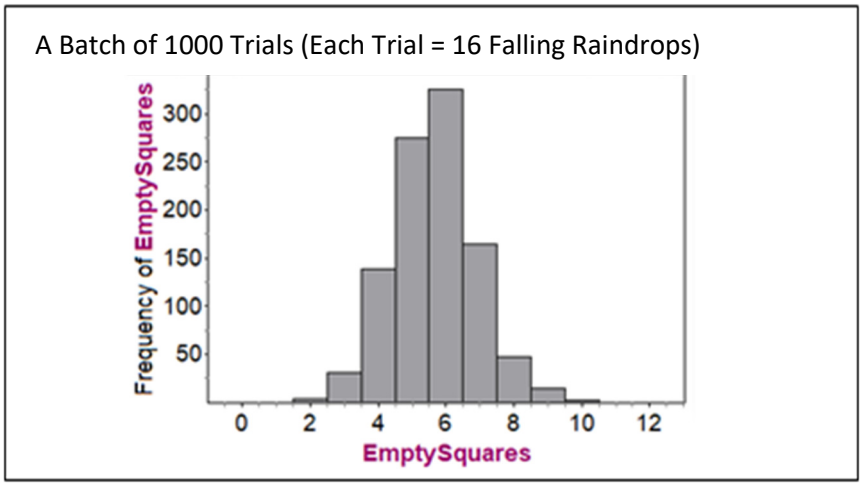


Figure 11: Counting the empty squares in each of 1000 trials

As the PSTs examined graphs of the larger sets of data for the numbers of empty squares for the Falling Raindrops task, it was clear that the graphs helped the PSTs see the emergence of a distribution, similar to that of Random Walks, which led them to look beyond just the measures of central tendency to focus on aspects of variability. For example, based on Figure 11 above, while students might typically expect around 5 or 6 empty squares (in a single trial of 16 falling raindrops), they would not be surprised by 3 or 4 empty squares, nor 7 or 8 empty squares. Yet they wondered about the true probability of "zero empty squares", which they knew would at least correspond to one raindrop on each of the 16 tiles – Precisely what so many of them predicted at the outset (and many even "centered" the raindrop within each tile).

In looking at repeated results from many trials, students also made a conjecture about variability in "the number of empty squares" as being related or correlated to "the maximum number of raindrops in any square". Again, this came only after seeing not only their own data from in-class experimentation, but from many results that were simulated by *Fathom*: The reasoning was that a high "maximum number" of raindrops in a single tile or square might be correlated with a higher number of empty squares. The computing power of *Fathom* makes conjectures such as these easier for students to investigate.

Lastly, beyond simply counting the number of empty squares, the issues of placement of the raindrops across the 4 x 4 grid seemed to be important for our students. For example, although "four empty squares" may not be surprising numerically, seeing those squares empty placed across the top row of the grid *would* be a surprise to some PSTs. In other words, the actual placement where empty squares occur should be random, according to some comments. Similarly, seeing four empty squares across the main diagonal of the grid doesn't suggest the kind of variability in placement that students were expecting. That is, students wondered how likely of an occurrence would a specific placement be?

CONCLUSION

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. In *Random Walks,* they found it easier to make probabilistic statements about results: If a batch of 35 games or trials all had results of more than 19, just as a hypothetical example, then PSTs would be inclined to say that was highly unlikely. In *Falling Raindrops,* students would similarly be suspicious a claim was made that a batch of 35 trials all had less than 3 empty squares in each trial. The likelihood of such events (35 random walks all having results more than 19, or 35 trials of falling raindrops all having less than 3 empty squares) is adjudged as quite low based on the variability they saw in repeated trials. Thus, experiencing variability in repeated trials made an influence on the PSTs' ability to make inferences.

In extending the issue of confidence in making inferences about a single trial, for the future we plan to ask students to respond to a scenario which posits a "Real or Fake?" aspect to a single trial of a *Random Walk.* We'll give our PSTs the following prompt:

You ask your class to go home and conduct a single trial of "Out of Bounds in 4", and bring back the resulting path the next day. One student hands in the graph below. You wonder: Did the student really do the trial, or just make up the results?

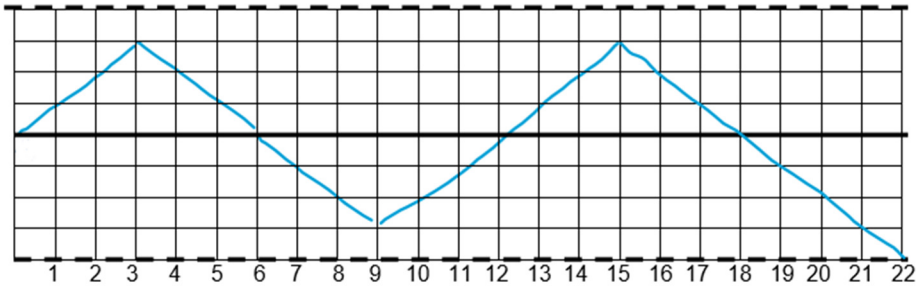
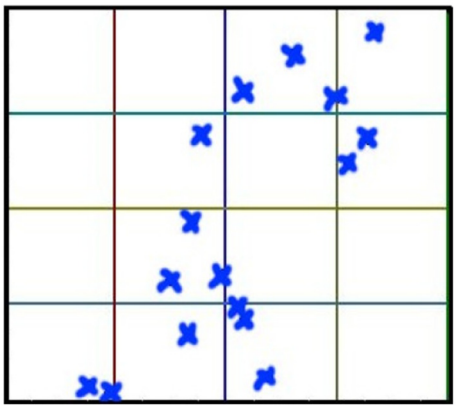


Figure 12: "Real or Fake?" A Question about a *Random Walk*

Our conjecture is that PSTs who've gone through the intervention described in this paper will debate the question: On the one hand, having a result of 22 in a single game (trial) of "Out of Bounds in 4" seems plausible. On the other hand, they'll likely want to review many trials to see if that exact pattern ever emerged (3 up, 6 down, 6 up, 7 down), and we imagine they'll question the probability of the "result" in Figure 12 happening naturally.

We did ask a related "Real or Fake?" question in the context of *Falling Raindrops*, and here is the prompt:

You ask your class to go home and conduct a single trial of "Falling Raindrops", and bring back the resulting grid the next day. One student hands in the graph shown.



You wonder: Did the student really do the trial, or just make up the results?

___ Those are real results
___ Those are made-up results
___ No one can have much confidence in whether the results are real or made-up

Figure 13: "Real or Fake?" A Question about *Falling Raindrops*

In Figure 13, we provided some prompts, and most participants decided that the results were made-up, with the next favorite response being that no one can have much confidence. Very few PSTs thought they were real results (which in fact they are, although we had to run *Fathom* through an extraordinarily high number of trials to obtain that exact placement). Again, while the total number of 7 empty squares did seem plausible, based on past results they had seen, most participants were skeptical of the "pattern" shown.

Our initial purpose in conducting the intervention using the two tasks as described was to informally gauge how the use of *Tinkerplots* and *Fathom* could bring aspects of variability to the forefront of PSTs' attention in probabilistic situations. 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 the software. A good example was when students asked a wait-time question concerning *Falling Raindrops*: "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 ≈ 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 (again, we haven't yet figured out how to replicate the *Falling Raindrops* scenario on *Tinkerplots*).

Perhaps the most intriguing question from students 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".

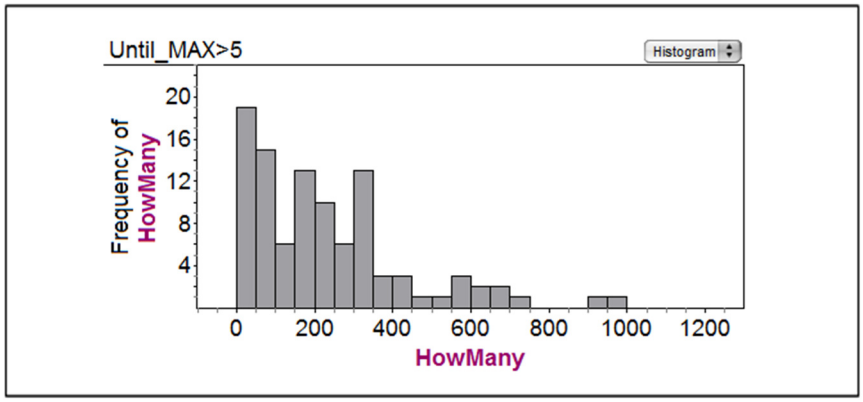


Figure 14: 100 Experiments of "How many trials to get a square with 6 or more drops?"

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 14.

Using the mean result from 100 experiments, which was about 230 trials for Figure 14, 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 on the results for both *Random Walks* and *Falling Raindrops*, 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 the two tasks. 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

Canada, D. (2006). Enhancing Elementary Preservice Teachers' Conceptions of Variation in a Probability Context. Statistics Education Research Journal 5(1), 36-63. [Online: www.stat.auckland.ac.nz/serj]

Cobb, G. & Moore, D. (1997). Mathematics, Statistics, and Teaching. American Mathematics Monthly, 104 (9), 801-824.

Davis, P. & Hersh, R. (1986). Descarte's Dream: The World According to Mathematics. San Diego, CA: Harcourt Brace Jovanovich.

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 (2005) 37: 16

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.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), On the Shoulders of Giants (pp. 95-137). Washington, DC: Academy Press.

Pfannkuch, M. (1997). Statistical thinking: One statistician's perspective In F. Bidduch & K. Carr (Eds.), Proceedings of the 20th Annual Conference of the Mathematics Education Research Group of Australasia , (pp. 406-413). Rotorua, NZ: MERGA.

Pfannkuch, M., & Wild, C. (1998). Statistical thinking and statistical practice: Themes gleaned from professional statisticians. Unpublished manuscript.

Pfannkuch, M., & Wild, C. (2001). What do we know about statistical thinking? Overview of statistical thinking, a literature review. In C. Reading (Ed.), Background Readings for the Second International Research Forum on Statistical Reasoning, Thinking, and Literacy, Armidale, Australia.

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

Reimers, S., Donkin, C., & Le Pelley, M.E. Perceptions of randomness in binary sequences: Normative, heuristic, or both? Cognition, Volume 172, 2018, Pages 11-25, [ISSN 0010-0277]

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Bidduch & K. Carr (Eds.), Proceedings of the 20th Annual Conference of the Mathematics Education Research Group of Australasia, (pp. 6-22). Rotorua, NZ: MERGA.

Shaughnessy, J. M. & Ciancetta, M. (2001). Conflict Between Students' Personal Theories and Actual Data: The Spectre of Variation. Paper presented at the Second Roundtable Conference on Research on Statistics Teaching and Learning, Armidale, New South Wales, Australia.

Watson, J., Kelly, B., Callingham, R. & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. The International Journal of Mathematical Education in Science and Technology. (34), 1-29.

Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. International Statistical Review, 67, 233-265.