Connecting Probability to Statistics using Simulated Phenomena

Theodosia Prodromou
theodosia.prodromou@une.edu.au
Department of Education
University of New England
Australia

Abstract: *This article addresses the use of probability to build models in computer-based simulations, through which exploring data and modelling with probability can be connected. The article investigates students' emerging reasoning about models, probability, and statistical concepts through an observation of grade 9 students, who used TinkerPlots2 to model a sample simulation based on probabilistic models of populations and tested models by comparing their behaviour with the generated data. Results from this research study suggest that students' use of probability to build models in computer-based simulations helps students to conceive of objects as comprising a set of data and the data distribution as being a choice made by the modeller to create approximations of real or imagined phenomena, where approximations depend on signal and variation.*

1. Introduction

The modelling approach as a means for bridging complex problems and school mathematics [1, 2], emphasizes the mathematization of real situations in a meaningful way for the learner [3]. The use of this approach involves Model Eliciting Activities (MEA) that provide students with an opportunity to mathematize a real world phenomenon and create models through repeated cycles of translation, description, and prediction of data [4].

The modelling approach reinforces uses of models that are formalised in a symbolic system and developed to represent concrete situations or problems arising from reality. Such modelling of concrete situations involves building models that are not always deterministic; models that incorporate uncertainty or random error in a formalized way are probabilistic models. These probabilistic models, according to their inherent rules, are expected to simulate the behaviour of random phenomena and also predict specific properties of random phenomena. For example, a physicist could define a sub-atomic phenomenon by making predictions about the probability distributions of various outcomes. Indeed, physicists often assign a probability distribution to the observable phenomenon under study rather than assigning a definite value. It is important to note that some quantum phenomena can only be described using probability functions.

The study of the modelling approach in learning sciences adapts operational definitions "by making predictions on the range of observable 'processes' that students will engage in when confronted by an authentic model-eliciting situation and the range of conceptual systems emerging from this engagement" (see [5], p. 130).

Many phenomena are not deterministic, so the teaching of probability is important. When teaching probability in schools, random generators can be seen as determined only if one is aware of a set of factors that causally affect the behaviour so that the outcomes are entirely predictable. In practice, this is an unlikely state of affairs and it is likely that one would be interested instead in adopting a probabilistic model.

In this framework, a probability distribution of some discernible characteristics has the status of a model of the data that describes what one could expect to see if many samples were collected from a population, enabling us to compare data from a real observation of this population with a theoretical distribution. This perspective is in accord with the *modelling* process of *contemporary* statistical thinking [6] that allows the application of theoretical results when making statistical inferences regarding particular observed sample statistics or data analysis. There is a growing body of research studies that reports how students used dynamic statistical software to generate models and simulations and answered informal statistical inference questions e.g. [7] and [8]. For example, Fitzallen and Watson (see [7]) research study showed that middle school students were able to represent data, create data summaries and make informal inferences when using *TinkerPlots* [9]. Maxara and Biehler (see [8]) studied college students' reasoning while creating models and running simulations when using Fathom software [10]. They documented that college students experienced difficulty in modelling statistical problems and simulating data. They noted that students continued to have certain probabilistic misconceptions although the outcome of a simulation provided evidence to the contrary.

This body of research studies does not only reveal some of the affordances of technology as a learning tool for teaching statistical concepts, but it also shows that probability is a central component for statistical investigations. This has engendered the need for using probability to deal with a variety of real-world situations that encompass interpretation of simulated data when teaching informal inference.

2. Probability as a Modelling Tool

Probability should not only be used to explain the methods of inferential statistics at the high school level but also as a tool for modelling computer-based action and for simulating real-world events and phenomena. It is necessary for curricula to "stress an alternative meaning for probability, one that is closer to how probability is used by statisticians in problem solving" (see [11], p. 1). Recently, international curricula have opted for a modelling approach [12] when teaching probability.

Recent developments in software provide the user with modelling tools based on probability (i.e., the user can set the probability of some event). These modelling tools could be used to construct probability models used by computer-based simulations.

One example of these recent developments in software is the *BasketBall* simulation developed and used for the doctoral thesis of Prodromou (see [13] and [14]). The *BasketBall* computer-based simulation simulated a basketball player attempting to make a basket. Underlying the simulation were two mechanisms for generating the trajectories of the balls following Newton's Laws of Motion; one was fully deterministic; the other used a probabilistic model that incorporated variation in the trajectories. The interface was designed in such a way that the data-centric perspective was presented graphically as a set of data about the trajectories and success of shots at the basket, and the modelling perspective was presented as the probability distribution that generated the varying trajectories of the balls.

Prodromou's research investigated how 15-year-old students, using the *BasketBall* simulation, co-ordinated the experimental outcomes (data-centric perspective) and the theoretical outcomes (modelling perspective) produced from a theoretical model [14]. The *BasketBall* simulation was inspired by the secondary curriculum that focuses on students' use of distributions to make inferences. It was anticipated that students might articulate a data-centric perspective on distribution that would be consistent with the Exploratory Data Analysis [15] approach, a means of engaging students in statistical analysis. One alternative view the students might adopt would be to recognise the probabilistic features of distribution that is the modelling perspective. Such a modelling perspective reflects the mind-set of statisticians when proposing models of how the data are generated out of random error and various effects. Statisticians often explain variation in data distributions as being either the result of noise or error randomly affecting the main effect, or higher random effects.

The second example is the *TinkerPlots* [9] a software that has been developed by a team (in the University of Massachusetts Amherst) led by Konold. The *TinkerPlots* software was designed to help students in grades 4-8 develop understandings of data, probability, and statistical concepts. *TinkerPlots* provides students with tools to “analyse data by creating colourful visual representations that will help students to make sense out of real data and recognize patterns as they unfold. Recently, a new version of *TinkerPlots*, *Tinkerplots2*, has become available and was used in this study. It offers new tools that exploit probability as a modelling tool using the sampler that is essentially a non-conventional form of probability distribution.

The student can select to sample using a mixer (top left in Figure 2.1), a spinner by defining the sizes of the sectors (top right), a histogram by determining the heights of the bars (bottom left), or a probability density function by drawing a curve to define a probability density function (bottom right). All these options of the sampler engine are elaborated further in the following investigation of grade 9 students’ use of this modelling tool to build a model of a real-world phenomenon, an exercise chosen because “the most effective way to create process knowledge is to develop a model that describes the behaviour of the process” (see [16], p. 230). This process knowledge, however, inevitably brings with it new challenges in how children learn and gives rise to research

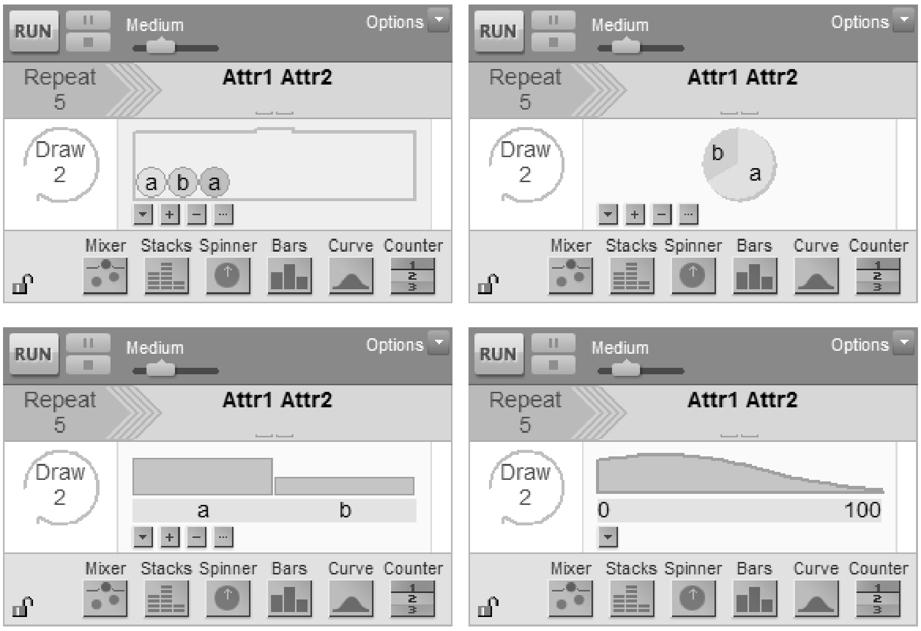


Figure 2.1 Examples of samplers in *Tinkerplots*2

questions about the conceptual development of students who will engage in connecting probability to statistics and to simulated real phenomena.

This article seeks to address the question of how the students are connecting probability to statistics when using simulated real phenomena.

3. Methodology: Data Collection and Analysis

The data used here come from an ongoing research study conducted in a secondary school. The data have been collected out of classroom. The students spent extensive time working in pairs. The researcher interacted continuously with the pairs of students in order to probe their reasoning and understanding. The data collected included audio recordings of each pair's voices and video recordings of the screen output on the computer activity using Camtasia software [17]. The researcher (Re) prompted students to use the mouse systematically to point to objects on the screen when they reasoned about computer-based phenomena in their attempt to explain their thinking.

At the first stage, the audio recordings were fully transcribed and screenshots were incorporated as necessary to make sense of the transcription. The most salient episodes of the activities were selected. The data were subsequently analysed using progressive focusing [18].

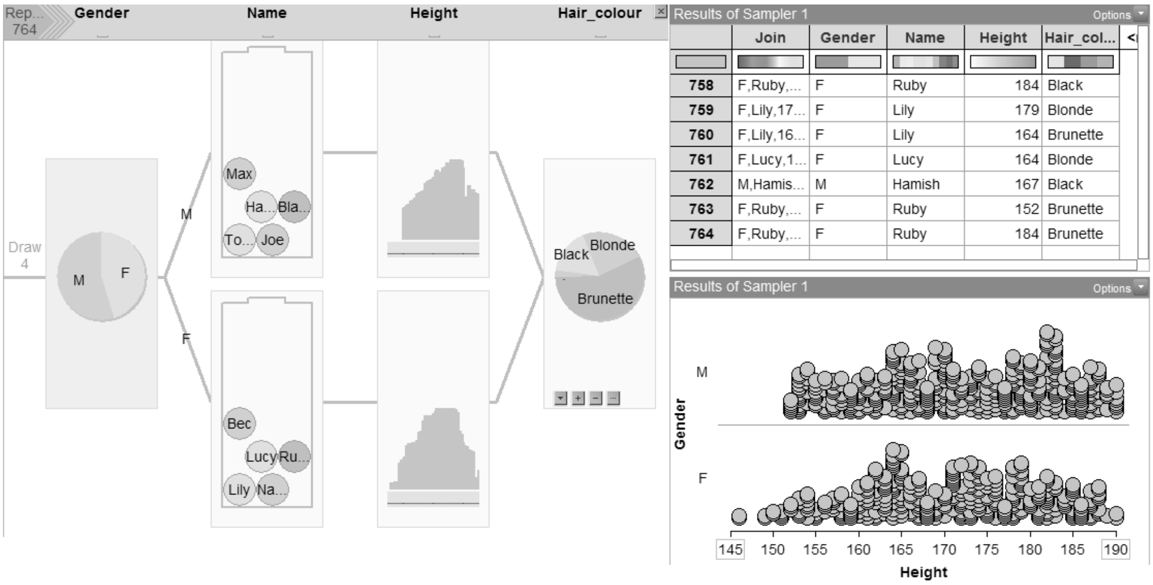
This article focuses on one pair of students, George and Rafael. Although the same insights as reported below were evident in the analysis of the sessions of other pairs of students, George and Rafael provided (in my view) the clearest illustration of how students build connections between probability and statistics, working with simulated phenomena within a *TinkerPlots2* computer-based environment.

4. Task

Students spent three lessons watching instructional movies that show how to use *TinkerPlots2* features to build simulations and spent three lessons (40 - 45 minutes each) building simulations. In the fourth lesson, students watched a *TinkerPlots2* movie that showed how to use *TinkerPlots2* features to build a data factory that simulates real phenomena. Students were shown how to set a sampler as a mixer, a spinner, a histogram, or a density function. As the simulation ran, the students observed the generation of data, and the distributions of the attributes of data. In the fifth lesson, students were asked to use the tools of *TinkerPlots*2 and the "Data Factory" feature of *TinkerPlots*2 (see Figure 5.1) to generate a number of individual "virtual students" to populate a "virtual secondary school," with each "virtual student" created using probability distributions for each of the different variables (gender, name, height) used to define them.

5. Results

To illustrate the students' work, I joined George ("Ge") and Rafael ("Ra") as they began to use the tools of *TinkerPlots*2 to create a factory for their virtual secondary school. They began by creating a sampler to assign gender to each virtual student. To do this they used a *TinkerPlots2*



	Join	Gender	Name	Height	Hair_col...
758	F,Ruby,...	F	Ruby	184	Black
759	F,Lily,17...	F	Lily	179	Blonde
760	F,Lily,16...	F	Lily	164	Brunette
761	F,Lucy,1...	F	Lucy	164	Blonde
762	M,Hamis...	M	Hamish	167	Black
763	F,Ruby,...	F	Ruby	152	Brunette
764	F,Ruby,...	F	Ruby	184	Brunette

Figure 5.1 Data factory that simulates a 'virtual school'

"spinner" (see Figure 5.1), in which they could visually assign different angles to correspond to the probability of a given outcome. As can be seen in Figure 5.1, they chose unequal angles for the sectors, thus giving unequal probabilities of getting male (m) students and female (f) students; in this case there was a much larger probability of getting male students than females. Students explained their choice:

1. Ge: In schools there are generally more males.
2. Ra: Umm, generally the schools that I've been in there have been more males. This school for example has a lot more males than females.

George and Rafael then chose to use the "mixer" function of *TinkerPlots*2 to give names to the virtual students. The boys created two different mixers, one for the names of each gender, and decided what names to place in each mixer from which one name would be chosen at random for each virtual student of the appropriate gender. When the boys went on to give the virtual students another characteristic, height, Rafael quickly stated, "This is getting complex. For a student, we have a choice of gender, name, and height. We are going to have many data for each student." George agreed saying that although "each student is one, varied information is provided for a student."

Rafael and George seemed to find it complex when a person or an object was presented as comprising a set of attributes. It was interesting how the boys used a modelling approach as an intermediate step in attempting to perceive a holistic entity, in this case a virtual student, as consisting of a cluster of pieces of data.

The second attribute the students decided to introduce for the students in the virtual school was height. The boys set up two samplers, one for boys' height and one for girl's height as a probability density function by drawing two curves as shown in Figure 5.1. They also set up a sampler for hair colour as a spinner, defining the sizes of its sectors.

Twenty minutes into the activity, the boys decided to make 675 virtual students. They ran the simulation to generate sample data for their virtual school and the researcher asked them to compare the distributions of the heights of the virtual students with the curves they drew in the sampler (see Figure 5.2).

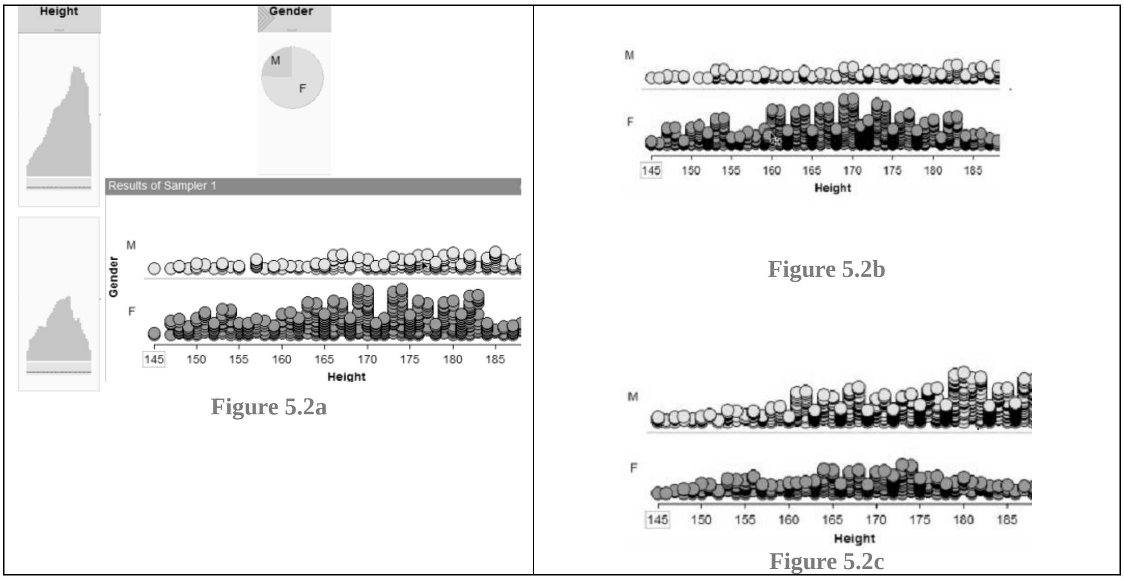


Figure 5.2a

Figure 5.2b

Figure 5.2c

Figure 5.2 Distributions created in the sampler and distributions of height

3. Ge: In schools there are generally more males.
4. Ra: Umm, generally the schools that I've been in there have been more males. This school for example has a lot more males than females.
5. Ra: (pointing to the distribution on the bottom right of Figure 5.2a). We had it rising there, then we had a drop. Then we had a big rise there which is why I'm guessing all these came from before it dropped down there.
6. Ge: Well the females model of the rise (pointing to the graph on the bottom left of Figure 5.2a), which is here (pointing to 150-155 of the bottom left graph), and that goes down a bit before continuing, so it would be going, down, continuing up (170-175), around here (180-183), before decreasing again. With the males, we thought the height was a lot higher than was there (pointing to the graph on top left).
7. Ra: There, there's a lot less males, though, so even the end results aren't as packed there; they're more spread out, or they seem to be spread out as much as the females but there's no piles. Like, you can't see them as high as well as you can see the females height increases, but you can actually see the increase there (175-185 right bottom).
8. Ge: There is an increase but it's not as prominent as the female increase.

Rafael and George then decided to generate 1000 virtual students because they believed that the graphs would show the distribution of the height of males more clearly.

9. Ra: But now you can see the increase in the males a lot better (pointing to the top Figure 5.2b). Like, you can see the towers higher, stacked higher than they were before. It's clear to see that there is increase, but with the females you can still see the increase, the huge increase the females have had as well.

Rafael paid attention to slices of prominent features of the distribution of males, such as higher areas of accumulated data compared to the behaviour of the male data in Figure 5.2a. He concentrated mainly on the distribution of males but he seemed to start developing an understanding of the impact that the selection of the gender by the spinner had on the distribution of males. He could not also dismiss the "huge increase the females have had."

Next, George attempted to equalise the number of male students with female students. After having 50:50 male to female students, the boys observed the new distribution (Figure 5.2c):

10. Ra: Yeah, it's a lot clearer to see the increase in the males (pointing to the distribution of male height) now; because it is a probability it's not going to look exactly like that (pointing to the

distribution of male height they created in the sampler). There's going to be exemptions. But you can see, you can see the overall that it's increasing. Getting higher here before dropping down again (pointing to the distribution of male height), which is what our graph showed (pointing to the distribution of male height they created in the sampler). The females you could tell fairly compact in the middle …and there's not as much any to the side which is shown here (pointing to the distribution of female height).

11. Ge: Exemptions?
12. Ra: I suppose there is always gonna be exceptions to the graph. They're not always gonna be exactly as we plan out. Because this is based on probability and probability, just because we see that (pointing to the distributions they created in the sampler), like it doesn't mean that it will follow that 100%.

The above excerpt shows that Raphael recognized the minimised impact of the variation introduced by the probabilistic selection of the gender variable. Rafael's attention was, at this point, focused on how the shape of the distributions of heights changed compared to the distributions they created in the sampler. He focused on the overall behaviour of the distribution of males' heights and females' heights and recognized that "there is going to be exemptions." In my view the word "exemption" was an expression attributed to the statistical concept of "uncertainty." The extended discussion of the boys showed that Rafael did not expect an absolute resemblance of the data distribution of heights to the distribution they created in the sampler due to uncertainty caused by probability.

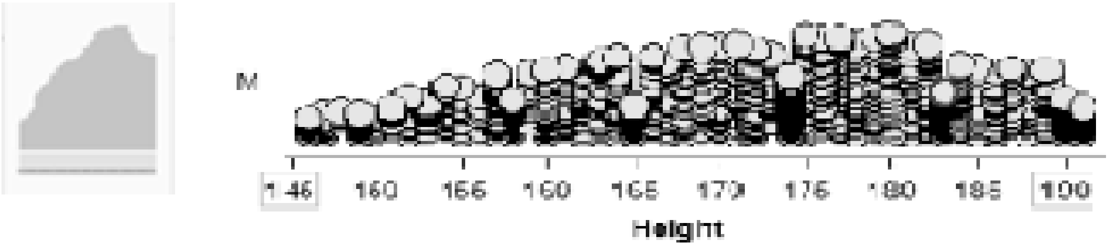


Figure 5.3 The distribution of male height they created in the sampler and the distribution of male height

In a later trial, George suggested creating a virtual school with 500 all male virtual students. To their surprise:

13. Ra: Well, it's steadily increasing with slight jumps. Like it lows there and jumps up a bit … guessing that 140, 155 region there (pointing to the left graph of Figure 5.3).
14. H: It's a 163 (left graph). And that 163 just there (right graph).
15. Ge: It's going up and just going down a little bit (right graph).
16. Ra: That's probably where that is come from 160. 164 here has just a low. But then it continues. Then it jumps back up and keeps going before dropping down a bit again here (pointing to both graphs at the same time).
17. Ge: Just there that one, for some reason it's just steadily dropping (pointing between 180-185 of right graph).
18. Ra: … except there (180-185). That seems to have a sharp drop, which I'm guessing is just off one of these areas here. Where it just seems to drop down (both graphs).
19. Re: Do you believe that the final graph resembles of what you created in the sampler?
20. Ra: Yeah, fairly accurately.

After excluding female students from their model, George and Rafael were more easily able to compare slices of the distribution of the observed data to the corresponding slices of data as represented on the density function that was drawn by them in the sampler. At the end, Rafael was able to appreciate that the two distributions eventually would accurately resemble each other.

6. Discussion

From our results I have described how Rafael and George moved over to the modelling approach by perceiving a holistic entity, such as a student, as consisting of a cluster of pieces of data having attributes such as gender, name, height, hair colour, etc. When the students attempted to simulate a virtual school through the activity of setting up the sampler as a spinner that defined the sizes of the sectors, they connected probability to the simulated phenomenon. Similarly, when the students set the sampler as a mixer and selected a sample randomly, they again drew connections between probability and the simulated phenomena. When they set a sampler as a probability density function by drawing a curve based on the users' personal experiences, they drew connections between probability, statistics, and the simulated phenomenon.

When the boys were challenged to draw a curve that mirrored a probability density function of heights, they first identified the area where the most common heights were accumulated (what we could consider the "signal," though they did not use this term) and then talked about the variation (what we could consider the "noise") of heights around that area. These students seemed to realise that the nature of a reasonable approximation of real or simulated phenomena lies in the relationship between signal and noise (lines 4, 11, 14, 16). A good co-ordination of signal and noise requires a fairly good knowledge of the phenomenon at hand, so that the students can design the Data Factory in such a way that it would generate sample data that would resemble as much as possible the real-world phenomena it was intended to model. These boys seemed to realise the importance of the choices they made as modellers, thus they focused their attention on the distribution of heights they created in the sampler and the other attributes, such as gender and hair colour. As we witnessed from the students' activity, the use of probability to build models in computer-based simulations, places distribution of data (e.g. distribution of students' heights) in the foreground debate not as a pre-determined entity but as a non-fixed entity, open to debate.

When the boys were working on the task of comparing the data distribution of the simulated heights to the distribution they created in the sampler, they appeared to have some difficulty understanding the nuances of probability. Nonetheless, they were not without insight. Rafael articulated situated heuristics--for example on line 10--that can be interpreted as a construction of a relatively naïve conception of the use of probability both as a modelling tool and as a measure of confidence that one can give to the model they created. However, another possible interpretation of this situation could be that when the model is run a few times, there is stability in the peaked data but there is some variation observed in the general details of the shape (line 8), thus Rafael could not recognize or accept the absolute resemblance of the distributions of heights to the curves the students created in the sampler (lines 19) due to the existence of many variables that caused variation. When they excluded one variable from the attributes, the students were better able to draw conclusions about the resemblance of the distribution of the generated data with what they designed in the sampler. By excluding variables from the model, thus reducing complexity, the students expressed an approach to introducing the variables in a systematic way that might benefit students exploring the connections between probability and the statistical data generated by simulated phenomena.

There is huge potential for future research in the area of technology in teaching probability. Curriculum supported by the use of dynamic software will become more prevalent in introducing probability as a tool for modelling computer-based action and for simulating real-world events and phenomena. This modelling approach of probability will give probability a place in mathematical curricula as a defined piece of mathematics. Observing students as they use probability as a modelling tool on computer-based simulations may provide deeper insight into how students connect modelling with probability to data and statistical concepts.

References

[1] English, L. D. (2008). Interdisciplinary problem solving: A focus on engineering experiences. In M. Goos, R. Brown, & K. Makar (Eds.), *Pro*ceedings *of the 31st Annual Conference of the Mathematics Education Research Group of Australasia*. (Vol.1, pp. 187-193), Adelaide: MERGA.

[2] Mousoulides, N., Sriraman, B., & Lesh, R. (2008). The Philosophy and Practicality of Modeling Involving Complex Systems. *The Philosophy of Mathematics Education Journal, 23*, 134-157.

[3] English, L. D., & Fox, J. L. (2005). Seventh-grader's mathematical modeling on completion of a three-year program. In P. Claarson et al. (Eds.), Building connections: Theory, research and practice (vol.1, pp. 321-328). Melbourne, Australia: Deakin University Press.

[4] Lesh, R. A., & Doerr, H.M. (2003). Beyond constructivism: Models and modeling perspectives in mathematics teaching, learning, and problem solving. Mahawah, NJ: Lawrence Erlbaum.

[5] Lesh, R. A., & Shiraman, B. (2010). Re-conceptualizing mathematics education as a design science. In B. Shiraman & L. English (Eds.), *Theories of mathematics education, advances in mathematics education* (pp. 123-149). New York, NY: Springer. doi: 10.1007/978-3-642-00742-2_41

[6] Wild, C., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*, 223-266. doi: 10.1111/j.1751-5823.1999.tb00442.x

[7] Fitzallen, N., & Watson, J. (2010). Developing statistical reasoning facilitated by TinkerPlots. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics* (ICOTS8, July, 2010), Ljublijana, Slovenia. Voorburg, The Netherlands: International Statistical Institute.

[8] Maxara, C., & Biehler, R. (2006). Students’ probabilistic simulation and modelling competence after a computer-intensive elementary course in statistics and probability. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education. Proceedings of the Seventh International Conference on Teaching Statistics* (ICOTS7, July 2006), Salvador, Brazil. Voorburg, The Netherlands: International Statistical Institute.

[9] *TinkerPlots: Dynamic data exploration* (Version 2.0) [Computer software]. Emeryville: CA: Key Curriculum Press.

[10] *Fathom: Dynamic Data*^{TM} *Software* (Version 2) [Computer software]. Emeryville: CA: Key Curriculum Press.

[11] Pratt, D. (August, 2011). *Re-connecting probability and reasoning about data in secondary school teaching*. Paper presented at 58th ISI World Statistics Congress. Dublin, Ireland.

[12] Chaput, M., Girard, J.-C., & Henry, M. (2011). Frequentist approach: Modelling and simulation in statistics and probability teaching. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics−Challenges for teaching and teacher education: A joint ICMI/IASE study: The 18th ICMI study* (pp. 85-95). New York, NY: Springer.

[13] Prodromou, T. (2008). *Connecting thinking about distribution*. (Unpublished doctoral dissertation). University of Warwick, UK.

[14] Prodromou, T. (2012). Students' construction of meanings about the co-ordination of the two epistemological perspectives on distribution. *International Journal of Statistics and Probability*, 1(2).

[15] Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison−Wesley Publishing Company.

[16] Hoerl, R.W., & Snee, R. D. (2001). *Statistical thinking: Improving business performance*. Pacific Grove, CA: Duxbury.

[17] Camtasia: Tec Smith Corporation (2010). *Catania studio* (Version 7.1) [Computer software]

[18] Robson, C. (1993). Real World Research. Oxford, UK: Blackwell.