YOUNG STUDENTS' WAYS OF CONSTRUCTING AND EVALUATING STATISTICAL MODELS

Sibel Kazak[1], Dave Pratt[2] and Rukiye Gökce[3]
[1]Pamukkale University, Turkey
[2]University College London, Institute of Education, UK
[3]Ministry of National Education, Turkey
skazak@pau.edu.tr

We report on research in which we aimed to develop young students' ideas about statistical models and modelling. In the study, students in a 6th grade classroom engaged in a data modelling task about jump lengths of two different paper frogs that involves formulating a statistical problem, identifying what and how to measure, deciding how to structure/represent data, comparing distributions and making predictions beyond experimental data. The data were analysed using progressive focussing. The study showed that most students tended to match "ups and downs" in the actual data when constructing their models. In evaluating models, students believed that the model data should look like the real data but had differing ideas about what made a good resemblance.

INTRODUCTION

Modelling and reasoning with statistical models are fundamental to statistical thinking during a data investigation (Wild & Pfannkuch, 1994). Through statistical models and modelling statisticians use data to predict and obtain information about the process generating data (Breiman, 2001). Existing research has offered some pedagogically effective ways of introducing this complex practice of modelling to young students in relation to the development of important statistical ideas, such as variation, signal and noise, uncertainty and inference.

Data modelling is such an approach in which learners construct (i.e. collect or generate) and use data to solve a statistical problem (Lehrer & Romberg, 1996). To examine students' thinking about natural variation, Lehrer and Schauble (2004) engaged 10–11 year olds in a data modelling task in which they investigated questions about plant growth by observing actual height change over time and under different conditions, e.g. fertilizers and light. When the question "what if we grow the plants again?" was posed, students made informal inferences and reasoned about uncertainty. Emphasis on distribution as a tool for thinking about the growth was also a critical part of the task. The findings indicated that, through generating, evaluating and revising student-invented displays or models of data, students developed an understanding of variation resulting from natural growth.

The study by English and Watson (2017) explored 11-year-old students' model generation for selecting swimming teams for the forthcoming Olympic games by using data on swimmers' performances (e.g., personal best times, number/level of competitions entered, ages etc.) in the previous competitions. This approach of modelling with data required students to interpret given data with a consideration of variability and problem context as well as to make informal inferences from their models. Students tended to acknowledge variation in a swimmer's performance in different events when referring to consistency in performance as an important variable to consider in model construction. Some groups also chose to compute the average of the variables to deal with the inconsistency in data. In their informal inferences, most groups expressed some uncertainty in selecting teams and applying their models to other sports events by referring to chance, which indicated their awareness of variability inherent in data. During model reporting in class, students also questioned each other's models especially when limited variables were considered in model construction.

The task based on purposeful computational modelling approach proposed by Ainley and Pratt (2017) engaged 11-year-old students in exploring the relationship between two or more variables within given small data sets and then building and revising models for generating data using *TinkerPlots* (Konold & Miller, 2011). The use of graphs in meaningful contexts and making predictions for imagined data helped students recognise signal within data, i.e., relatively stable aggregate properties of distributions. Students also tended to express noise (due to chance

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

variability) by using a range of values in the process of building a model in *TinkerPlots*. When evaluating the effectiveness of the model in generating the required outcomes, students used graphs to compare the model data and original data. This revealed different kinds of criteria for model evaluation: (1) assessment of whether the model was working as expected by comparing generated outcomes to the structure of the model built in *TinkerPlots*; (2) the resemblance between the model data and original data by comparing the simulation output with the original data; (3) assessment of whether the model produced realistic data, or impossible outcomes.

Drawing upon the earlier studies, in this paper we aimed to explore young students' construction and evaluation of statistical models through modelling with data during a task that required them to collect and represent data, interpret and compare distributions and make predictions beyond experimental data. The following research questions are considered:

- How did students model empirical data distributions based on their experiments?
- What criteria did they use for evaluating those models?

METHOD

Adopting a design study method (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003) we conducted a teaching experiment in a 6th grade classroom in spring 2017. The school was a large urban middle school (with approximate enrolment of 1600 students from 5th to 8th grades) located in Denizli, Turkey. 30 students (13 boys, 17 girls) of ages 11–12 took part in the study. At the 6th grade students were introduced to formulating research questions, collecting data, making frequency tables and bar graphs (no dot plots), computing and interpreting the mean and the range of a data set and using them to compare two data sets in mathematics. They had also prior experience of conducting experiments that involves taking measurements and recording data in science classes. However, they had no experience of modelling and using computer data analysis tools, such as *TinkerPlots*.

In collaboration with the mathematics teacher (the third author of this paper) of the 6th grade classroom, we designed and implemented a data modelling task, called Frog Olympics, over 9 class periods. During the task students worked in groups of 4–5 students and the role of the teacher and researcher was to facilitate group work and class discussions. After introducing the 100-meter jumping race rules in the Frog Olympics, students were given two different frogs made by origami (Figure 1) and asked to determine which of the given two frog designs to choose for 100-meter 'jumping' race.

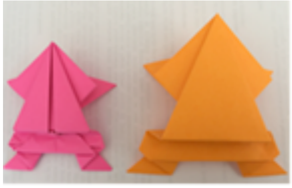


Figure 1. Two different frog designs used in the Frog Olympics task (pink frog is smaller than orange one)

The task structure included the following stages: 1) Introducing the game and planning how to choose between two frog designs; 2) Planning the experiment and collecting data; 3) Representing data; 4) Analysing data and making inferences; 5) Sketching a model for prediction; and 6) Evaluating the models. While the first five stages of the task were conducted in the classroom, the last stage of evaluating models (on which we place some emphasis in this paper) was done through 20-30 minutes long interviews with selected groups for more in-depth analysis of students' thinking.

To examine how students construct and evaluate models of distributions, we combined the experimental data from all groups in the dot plots for each frog design. The combined data distributions (Figure 2) were displayed on the classroom interactive board and students were asked to predict what the two frogs might do in many repeated jumps in the future using the following scenario: "A mobile game developer wants to make a digital version of each frog design for a game. Your task is to help the developer using the data you collected from flipping paper frogs. By looking at the dot plots of jump distances of pink and orange frogs, as a group predict how the distributions of jump distances of each frog design would likely to be if we were to collect more data." We introduced sketching as a way to generate a model of expected results before using *TinkerPlots* for modelling. Each group sketched their prediction on a given empty horizontal axis for the jump lengths scaled from 0 to 100 for each frog design on their worksheet.

During the interviews, each group was asked to evaluate the given 3-4 models (including theirs) and rate them as either "good", "OK", "bad" with reasons. Additionally, the interviewer showed the students the *TinkerPlots* model built based on their hand drawn sketch. After the interviewer explained how the *TinkerPlots* model was built and how the simulation works, they were asked to describe what would happen when they ran the simulation. Next the simulation was run and students were asked about whether the results turned out the way they expected. After testing their model with several simulations, they were asked to re-evaluate the given models.

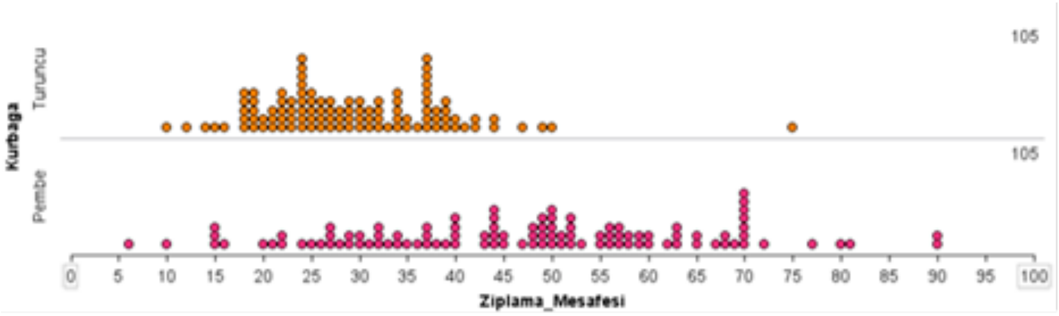


Figure 2. Dot plots of combined data showing the distance jumped on the horizontal axis for orange frog (at the top) and pink frog (at the bottom)

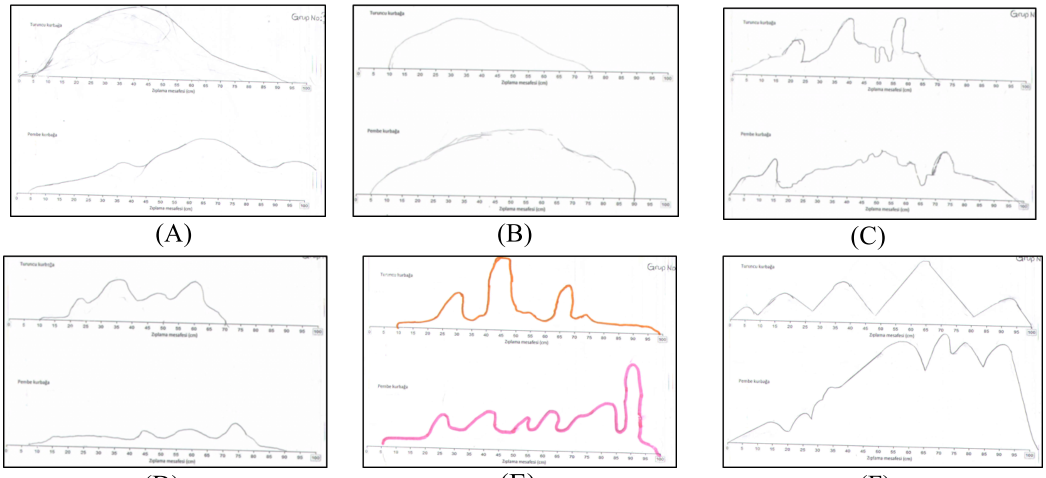
In this paper we will briefly describe students' work prior to sketching their models for prediction and mainly focus on how students modelled empirical data distributions based on their experiments and what criteria they tended to use for evaluating those models. Thus, the qualitative analysis of documents including written artefacts from each group work and transcripts of audio recordings of group interviews has used progressive focussing (Parlett & Hamilton, 1977) to describe and interpret students' ways of modelling distributions and evaluating various models.

FINDINGS

In the early stages of the task students had hands-on experience of collecting data using two different origami frogs. To decide which one of them to choose for the Frog Olympics, all groups displayed their data using a bar graph for each frog type. When comparing the results from the bar graphs, students tended to express variability in the data with reference to regularity and consistency of each frog's jump lengths as seen in English and Watson (2017). They also chose to calculate mean and range of each data set to make a decision. Then they were prompted to make dot plots of their data in order to shift their focus to the distributional features of data, such as shape, central tendency and variability. This helped them to see data as a distribution while they were comparing the jump lengths of two frogs by focussing on where the data were clustered and how spread-out they were. Perceiving data as a distribution of values, which was also emphasised in earlier studies (English & Watson, 2017; Lehrer & Schauble, 2004), helped with a natural transition to introducing sketching as a way to predict the results in the long run in the next stage of the task.

Model Construction

From our analysis of sketches drawn by the groups in the class and students' explanations of their model during the interviews, we identified various tendencies used when constructing a model based on empirical data (see Figure 2). As seen in Figure 3, most groups tried to match "ups and downs" in the actual data but not the jump lengths on the horizontal value axis. Therefore some of the models started from the same value as the real data but ended beyond the maximum value in the empirical results. Two groups actually matched the minimum and maximum data values in their models while others tended to construct their model a bit lower and higher than the actual range of data. The curves in the models also tended to run a little higher than the maximum heights of the actual clusters of data. Two distinct categories of models emerged from these sketches. As seen in the examples of models created by different groups (Figure 3), students mostly showed several small ranges of jump lengths in the data, which led to a series of 'ups and downs' (models C, D, E and F). A few other groups tended to base their idea of a modal clump around a broad range of jump lengths (models A and B) as they were using a central range of values for what seemed to be typical (Konold, Robinson, Khalil, Pollatsek, Well, Wing et al., 2002). Next we will elaborate on what criteria they tended to use when evaluating different models during the interviews by focussing on two groups.



(A) (B) (C)

(D) (E) (F)

Figure 3. Models constructed by the groups who were interviewed (top sketch for the orange frog and bottom sketch for the pink frog in each model)

Model Evaluation Criteria

Group 2 generated one of the models that showed the general trend (model A in Figure 3). This group determined a modal range of values where the most jump lengths were, such as 10-65 for the orange frog, and made the curve higher in that range. In the model data they also tended to go lower and higher than the minimum and maximum values of the actual jump lengths. In the interview, the group was asked to evaluate the models A, D, E and F as "bad", "OK" or "good". After an examination of each model, students reached an agreement on their evaluations: "good" for model E, "OK" for model A and "bad" for models D and F. They reasoned that the model E was "well thought out" and looked like the real data. Their model (A) on the other hand was considered as a "rough sketch" even though they believed it represented the most common jump lengths. When judging models D and F, students focussed on the maximum values of jump lengths for the model (70 for the orange frog, 90 for the pink frog) in model D and the disproportionately high curve between 85 and 100 for the pink frog in model F.

After watching the simulation of their model built in *TinkerPlots* (Figure 4) several times, the group was satisfied with the results from a large number of trials. When they were asked to evaluate the given models again, they wondered how the results of model E would look like. Since this was the first time students were introduced to *TinkerPlots* modelling tools, they were not sure if the other model would yield similar results to theirs. To test this, the interviewer ran the simulation of model E in *TinkerPlots* (Figure 5) a few times. Seeing this model data led students to change their initial evaluations to "good" for model A (theirs) and "OK" for model E because they thought that the simulation results had a series of squiggles for the orange frog while there was a bigger cluster of jump lengths in the real data.

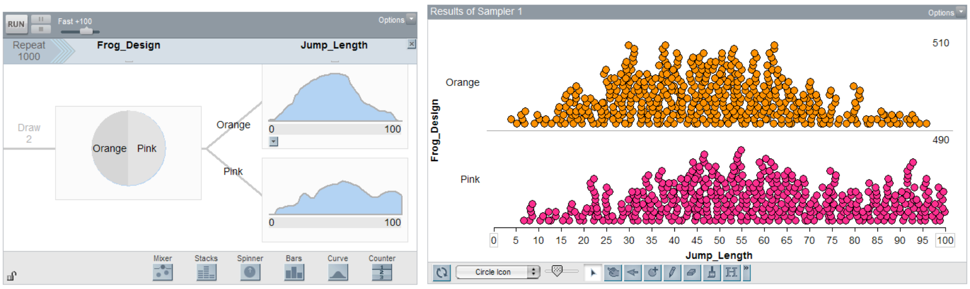


Figure 4. On the left the Sampler built for 'model A' in *TinkerPlots* using the curve device and on the right the simulation results from a large number of trials

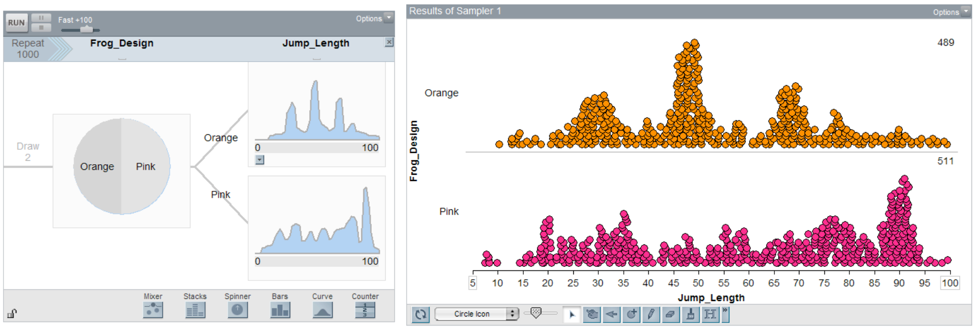


Figure 5. 'Model E' created in *TinkerPlots* using the curve device and the simulation results

In contrast to Group 2, Group 5 made one of the models showing several small ranges of jump lengths in the data (model D in Figure 3). In their model construction, this group paid attention to the minimum and maximum data values as well as tendencies to cluster or spread in the real data. During the interview, the group was asked to assess models A, B, C and D as "bad", "OK" or "good". After discussing each model's resemblance to the real data, students agreed that their model (D) and model C were "good"; models A and B were "bad". They initially argued that models C and D seemed "more thorough" while models A and B looked "sloppy". Then they did a match test between each model and the actual data distributions by superimposing the sheet with the model onto the sheet with the empirical data distributions as seen in Figure 2. They paid attention to matching "ups and downs" in the model and the real data. This test allowed them to confirm their initial evaluation of models. After seeing the simulation results of their model built in *TinkerPlots* from several runs, students concluded that the simulation results were as good as they expected.

In evaluating various models, the other groups we interviewed had similar tendencies to the ones reported here. For example, the group that created model C made their evaluations based on the match test between each model and the actual data used by group 5 while the other group that made model E paid more attention to the start and end values of the models rather than the "ups and downs" in their evaluations. Similar to the findings of Ainley and Pratt (2017), all groups in general expected that the model data should look like the real data but had differing ideas about what made a good resemblance.

CONCLUSIONS

There has been great interest in teaching statistics through modelling starting from young ages in the context of developing important statistical ideas, such as variation, signal and noise, uncertainty and inference (e.g., Ainley & Pratt, 2017; English & Watson, 2017; Lehrer & Schauble, 2004). The findings of this study, though based on limited data, supports the notion that such approaches have the potential to develop students' understanding of key statistical ideas and procedures as they construct a model. When students sketched their models, a consideration of central tendency, variability and uncertainty was apparent in various cases. Moreover, the use of *TinkerPlots* to run several simulations of student-generated models helped some students take their attention away from the "ups and downs" in the data and see that their own general trend curve looked better for making predictions beyond real data. However, this paper also indicates some of the complexity in the idea of modelling, as students will tend to focus on individual clumps or data values (minimum and maximum) in the real data rather than the properties of the aggregate as a whole when constructing and evaluating models of data. Thus young students would need more experiences for developing an intuition for seeing an overall shape of the distribution rather than a pattern of “ups and downs” when the amount of data gets larger.

ACKNOWLEDGMENT

The first author acknowledges the support of PAU-ADEP in the form of a conference travel grant funded by 2018KRM002-272.

REFERENCES

Ainley, J., & Pratt, D. (2017). Computational modelling and children’s expressions of signal and noise. *Statistics Education Research Journal, 16(2), 15-37.*

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science, 16*(3), 199-215.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Shauble, L. (2003). *Design experiments in educational research. Educational Researcher, 32(1)*, 9-13.

English, L. D., & Watson, J. (2017). Modelling with authentic data in sixth grade. *ZDM Mathematics Education*. DOI 10.1007/s11858-017-0896-y

Konold, C., Robinson, A., Khalil, K., Pollatsek, A., Well, A. D., Wing, R., & Mayr, S. (2002). Students' use of modal clumps to summarize data. In B. Phillips (Ed.), *Proceedings of the sixth international conference on the teaching of statistics* (ICOTS-6), Cape Town, South Africa, [CD-ROM]. Voorburg, The Netherlands: International Statistical Institute.

Konold, C., & Miller, C. D. (2011). *TinkerPlots 2.0: Dynamic data exploration*. Emeryville, CA: Key Curriculum.

Lehrer, R., & Romberg, T. (1996). Exploring children’s data modeling. *Cognition and Instruction, 14*(1), 69-108.

Lehrer, R., & Schauble, L. (2004). Modeling variation through distribution. *American Education Research Journal, 41*(3), 635-679.

Parlett, M., & Hamilton, D. (1972). *Evaluation as illumination: A new approach to the study of innovatory programs, Occasional Paper no 9*. University of Edinburgh. Edinburgh: Centre for Research in the Educational Sciences. Retrieved from ERIC database (ED167634).

Wild, C. J., & Pfannkuch, M. (1999). *Statistical thinking in empirical inquiry. International Statistical Review, 67*(3), 223-248.