Model-based Informal Inference

Theodosia Prodromou[1]

[1] University of New England, Armidale, Australia

Correspondence: Theodosia Prodromou, University of New England, Armidale, Australia.
E-mail: tprodrom@une.edu.au

Abstract

Following recent scholarly interest in teaching informal linear regression models, this study looks at teachers' reasoning about informal lines of best fit and their role in pedagogy. The case results presented in this journal paper provide insights into the reasoning used when developing a simple informal linear model to best fit the available data. This study also suggests potential in specific aspects of bidirectional modelling to help foster the development of robust knowledge of the logic of inference for those investigating and coordinating relations between models developed during modelling exercises and informal inferences based on these models. These insights can inform refinement of instructional practices using simple linear models to support students' learning of statistical inference, both formal and informal.

1. Introduction

Mathematical models are used in Statistics to represent a general pattern of data (Moore, 1990). In statistical research, modelling is central. Nonetheless, covariate adjustment (Cochran, 1977), a classical technique, could reduce the variance of the estimate without the need of the model to be precisely specified (Lin, 2013; Lu, 2016).

Statistical modelling has typically been a topic addressed in advanced mathematics courses, although its use from kindergarten through introductory statistics has been an emerging development (English, 2012; Lehrer, Kim & Jones, 2011). Lehrer et al. (2011) characterise statistical modelling: "Data modelling integrates inquiry, the generation of data, chance and inference". The practice of statistics is a form of modelling, as the development of models of data, variability, and chance are paving the way of the statistical investigation (Lehrer, Kim, Ayers & Wilson, 2014; Wild & Phannkuch, 1999). The aim of statistical practice is to make inferences from data. Making these inferences formally is nevertheless difficult for learners because it involves notoriously challenging concepts such as the process of hypothesis testing, probability density functions, and theoretical distributions (e.g., binomial, Poisson, normal, etc.). Informal approaches to statistical inference eschew these formalisms and introduce students to core ideas and techniques in ways that are accessible to learners who are not familiar with formal procedures, thus laying a foundation for their future study of formal statistical inference (Pratt & Ainley, 2008). If informal inferences can provide such a foundation for students' future study of statistics, we must ask questions about how to accomplish this aim in the classroom (Pfannkuch, 2011). My intent hence is to seek literature that aimed for a broad interpretation of modelling that would capture informal statistical inference and the generation of data through activities involving uncertainty. Research studies use modelling as a particular approach to informal statistical inference focusing on issues of modelling chance that are relatively accessible to students (Lehrer et al., 2014; Prodromou, 2012). These studies focused on engaging students in building and revising models of chance processes that either generate repeated measures of the same attribute, e.g., the angle at which a basketball is thrown in the basket (Prodromou & Pratt, 2006), or by producing a product, e.g., producing bricks. The longer progression of learning that situates this form of model-based reasoning (Lehrer & Shauble, 2010) that "entails deliberately turning attention away from the object of study to construct a representation that stands in for a phenomenon by encapsulating and enhancing its theoretically important objects and relations. Instead of directly studying the world, one studies the model –the simplified, striped down analog". Lehrer et al. (2014) designed a pedagogical approach, intended to lead young students through the following steps in a trajectory to culminate in model-based informal inference, as involving the following steps:

1. Create a representation of the results derived from a repeated measures process.

2. Compare visualisations of the resulting data to foster an image of the shape of the data by choosing structure for the data (e.g., ordering cases, creating intervals).

3. Compare estimates of a true measure during the measurement process or a target value during the production process. These comparisons constitute statistics as measures of signal and noise (Konold & Pollatsek, 2002).

4. Measure uncertainty in the behaviour of random devices, such as dice or spinners. The way that they measure uncertainty develops informal seeds to include sampling distribution.

5. Reflect on the processes of repeated measures by analysing different sources of error that are modelled and represented as a combination of signal (sample average or median) and noise.

6. Investigate model fit, using informal criteria such as approximations of the shape of the data, as well as its center and spread.

7. Make informal inferences about the new sample statistics based on their belief about the adequacy of their model. The informal inferences are made using the sampling distributions of model parameters as a guide to probabilistic inference.

This learning trajectory of model-based inference necessitates coordination among practices and basic concepts (Lehrer et al., 2014), such as average, spread of data around the average, number of data cases in each bar, variability of the data, while they reason about the spread of data across a range of values, from which a trend of data or model can be "discovered". The above trajectory of learning is a theoretical description of student behaviour, but when considering this trajectory of inference, I have been curious whether it can help students develop an appreciation of the robustness or power of their inferences without constructing a "modelling" perspective alongside their "data-centric" perspective (Prodromou & Pratt, 2006). In Prodromou and Pratt, 'modelling' is defined as using a model—theoretical model of data—that generates the data, and this is similar to the conventional definition of statistical modeling during which a statistical model that embodies a set of assumptions concerning the data-generating process, generates the data. The assumptions embodied by a statistical model describe a set of probability distributions, some of which are assumed to adequately approximate the distribution from which a particular data set is sampled. These probability distributions that are inherent in statistical models are what distinguish statistical models from other, non-statistical, mathematical models. Whereas, the data-centric perspective moves from data to model/function (Engel & Kuntze; 2011, Tukey, 1977). The emphasis is on data, and the modellers begin the modelling process with the study of the data and their distinctive behaviour, (e.g., trend of data, pattern, and variation) in order to build a model or a function. One focal question of interest to this chapter: How is model-based informal inference/reasoning actually generated and what kind of learning arises from this bidirectional modelling process?

2. Theoretical Framework

In the literature, theoretical attempts to describe the modelling process and research studies that report on what happens when an individual is involved in mathematical modelling both typically include one or more of the following components: (a) the phases of the modelling process, (b) the required mathematical competencies, and (c) the interplay between these components. There are several frameworks describing the process of modelling a mathematical problem and these frameworks are often represented in a diagrammatic form (Blomhøj & Jensen, 2007; Blum & Leiß, 2007; Galbraith, Stillman & Brown, 2010). Such representations are illustrative of the different actions undertaken during the mathematical modelling process and the kinds of cognitive and mental activity that modellers engage in as they follow different steps of the mathematical modelling process. In what follows, we focus on one depiction of the process of bidirectional modelling for fostering statistical modelling that incorporates the building of connections between real data contexts, data, and statistical models. Before we will invest in further explanations of this diagram of statistical modelling, let us consider the background theoretical framework that underlies the development of such a theory of bidirectional modelling. The proposed diagram derived from thinking about statistical modelling as incorporating two distinct processes that have different beginnings. The one process begins with data which gives birth to the model or function, whereas and the second process focuses on the model that generates the data. Now, let us take a closer look at the crucial aspects of what might be termed as "ideal behaviour", in which modellers proceed effortlessly from data to model and vice versa during the statistical modelling process, and the learning that arises from the different stages of the process of model-based inference as depicted in the bidirectional modelling diagram (Fig. 1). In the "real data" branch of the framework, the first component involves posing questions about data or particular aspects of data. For example, posing questions about which kind of data would better answer a question within a context. The second component, collecting data, includes a variety of data collection or generation techniques like population sampling or probabilistic computer simulations. The third component involves displaying the data, and a careful observation of the distribution of data in order to extract important variables, detect outliers and anomalies, and form hypotheses worth testing, seeking common causes of the pattern and for particular causes in the deviations that might sometimes shed light on the problem situated within the context. The fourth component involves context-sensitive decision making about the question based on experimental data, and making inferences about the probability distribution that best describes the observed data. The latter could give interesting insight into the context. The fifth component, evaluating the model, involves evaluating whether the probability distribution or the linear regression model describes the data and comparing the behaviour of the

model to observed data. If the model fully explains the data, then the decisions about the model and characteristics of data are communicated, justified and reported. If modellers revise the model and the model does not satisfactorily explain the data, then they return to the third component of the "real data" branch of the multidirectional modelling process, to seek real data contexts, data, and probability distributions. Another model is built that better explains the observed data and if the model is deemed satisfactory, modellers communicate, justify and report on that model.
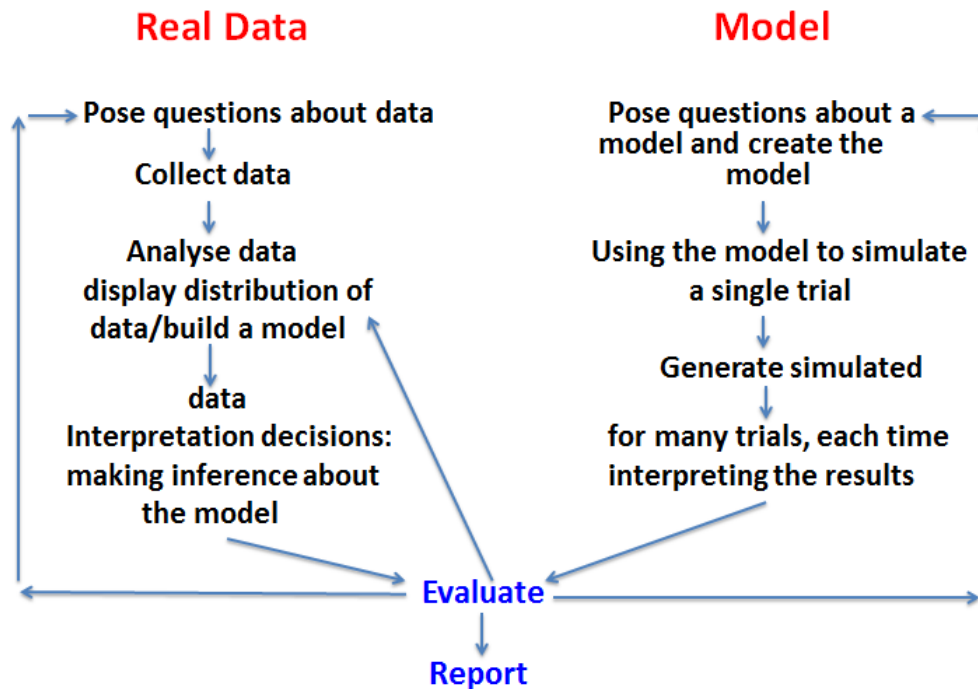

Fig. 1. Bidirectional modelling diagram of model-based informal inference

In the other direction– the "model" branch of the Bidirectional modelling diagram of model-based inference–, it is assumed that there is an unknown stochastic model that generates data. This model could be used either for prediction of the response variables to the future input variables, or for extracting information about the relationship between the response variable and the input variable. For example, the line of best fit in linear regression, logistic regression, or logit regression is a type of probabilistic statistical classification model. The trajectory of learning that is described by the model branch involves the creation of an informal model that generates data whose distribution mimics real-world samples. The modellers pose questions about a model that will generate data to simulate the problem; they investigate the target situation that is planned to be modelled, and they construct an informal model. Such models could be made by using software applications, such as the sampler from *TinkerPlots2,* which provides tools to construct models, for example, by describing a histogram by determining the heights of each bar, or by drawing a curve to define a probability density function. This first component involves building a model that relies on signal, variation, and spread of data. The second component involves using the model to generate a single trial of the simulation, investigating the outcomes from a single trial that involves one whole set of data, constructing an appropriate representation of the data from the single trial. The second component involves analyzing the data from a single trial and considering possible outliers and other individual cases. The third component involves using the model to generate simulated data for many trials, each time interpreting the results of the simulated outcomes while using the distribution of the simulated data to assess particular outcomes. The fourth component, evaluating, involves informal evaluation of the model by comparing the behavior of the simulated data (center of data, spread of data, variation of data, shape of the distribution of data) to the behavior of the real data, using student-invented criteria such as informal comparison of samples of simulated data to the expected probability distribution defined in the model. Evaluation of the model also involves comparisons amongst the center of the simulated data and the centre or average of the probability distribution; and comparisons amongst simulated spread of the data and the spread of the probability model. Such a comparison may prompt modellers to make interpretative decisions about whether the model fully explains the data. Then a modeller either decides to complete the model-based inference process and the decisions about the model and characteristics of data are communicated, justified and reported; or the modeller decides to change the model and repeat the phases of the modeling process in the "model" branch of the Bidirectional modelling diagram.

A recent study by Lehrer et al. (2014) conjectures that processes involving signal and noise are apt points for entry of informal inference in the process of building and refining models of chance processes due to two reasons. First, these processes are often tangible and therefore afford student opportunities to coordinate the process with the distribution of outcomes. Second, the process provides opportunities to perform an analysis of the source of variability and to model the variability of each source. Although young modellers of that study showed an understanding of sampling distribution of model parameters, they had not yet mastered the intricate coordination among chance, models, and inference. Understanding how students master model-based informal inferences has great potential to inform the field of statistics education (and other fields). Hence, the types of reasoning and conceptual structures that arise for students during their engagement with the process of model-based informal inference become an important feature of research. However, such research would be poor without ensuring that we have teachers who can support and strengthen students' model-based informal reasoning. In this light, I acknowledge the increased need for research on the knowledge that teachers have about model-based inference when they engaged in the process of a bidirectional informal model-based inference. Hence, this research study researches the knowledge that teachers have about model-based inference. To provide a finer-grained portrait of teachers' understandings of model-based inference, I conducted a design experiment (Cobb et al., 2003) involving 11 pre-service teachers and 6 in-service teachers. The design experiment instantiated the learning described as arising from the different stages of the bidirectional modelling diagram of model-based informal inference.

This study's engagement with model-based informal inference asked teachers to fit a line to data by eye, replacing formal algebraic construction of models with an informal method of generation.

3. Method

I selected a convenience sample of 11 pre-service teachers and 6 in-service teachers. Two flexible task-based interviews were administered individually to each participant. These interviews served as the primary source of data. Each interview lasted from 20 to 25 minutes. All interviews were videotaped. The flexible task-based interviews were composed of tasks intended to elicit teachers' knowledge about model-based inference regarding the following aspects of the line of best fit and its features: (a) develop the line of best fit (model) after plotting all the data (first interview), (b) criteria for evaluating the goodness of fit of the model composed as fixed signal and random noise (first interview), (c) revising line of best fit to produce different linear models and discuss the criteria guiding their informal inference (first interview and second interview), (d) discuss criteria guiding informal inference in a framework of letting the model generate data and evaluating the goodness of fit of the model (second interview). The data sets used in the tasks (a-b) were chosen so that a linear model would be appropriate for each of the tasks. In addition, none of the data presented had outliers or influential points to distract the participants. To assist and illuminate our analysis we shall begin by considering the following two concrete tasks undertaken by the teachers. The task undertaken during the first interview was related to "real world data" branch of the bidirectional modelling diagram of model-based informal inference. The tasks undertaken during the second interview were related to "model" branch of the bidirectional modelling diagram of model-based informal inference.

The first task asked teachers to investigate how an alligator grow. Secondary data were sourced from www.SteveSpangler.com. The teachers graphed the data of the measurements of the length of an alligator for 107 hours in a scatter plot (Fig. 2). The teachers were asked to determine the

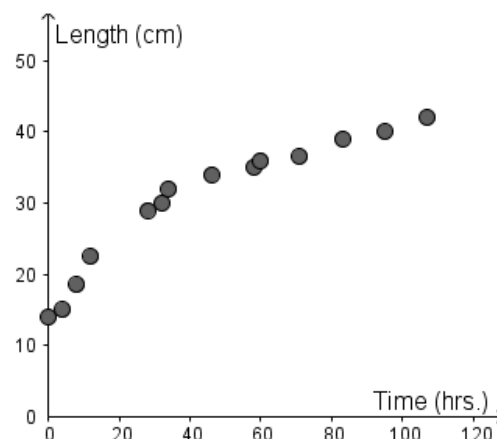| Amount of time from initial measurement (hrs.)	Length
0	14
4	15
8	18.5
12	22.5
28	29
32	30
34	32
46	34
58	35
60	36
71	36.5
83	39
95	40
107	42

Fig. 2. Data table, scatterplot

line of best fit for all the data points. The intention was to probe teachers' understanding of the informal regression line of best fit and its association with the variability of the data given the context. Teachers viewed a regression linear model displayed in a scatter graph together with the data points of the amount time from initial measurement and the length of the alligator. The teachers used GeoGebra software (Fig. 3) to build a linear model and were asked to judge the adequacy of the linear model in light of the 14 data points. The intention of this task was to elicit teachers' criteria of line of best fit and the extent to which data variability was coordinated with the line of best fit. They discussed the criteria guiding their informal inference and they calculated manually (Fig. 3) the slope of the line best fit and the equation of the line of best fit. The intention of this task was to probe teachers' understanding of estimating reasonably the location of the linear trend and assessing which criterion was more important to determine an informal line of best fit.

The second task required teachers to observe the generation of data by a linear regression model that relates the height, and the weight of humans (Fig. 4). They were asked to describe what the

Using data point (12, 22.5) and (95, 40)

$$m = slope = \frac{40-22.5}{95-12} = \frac{17.5}{83} \approx 0.21$$

Using (95, 40)

$$y = m \cdot x + b = 0.21 \cdot (95) + b$$
$$40 = 0.21 \cdot (95) + b$$
$$40 = 19.95 + b$$
$$b = 20.05$$
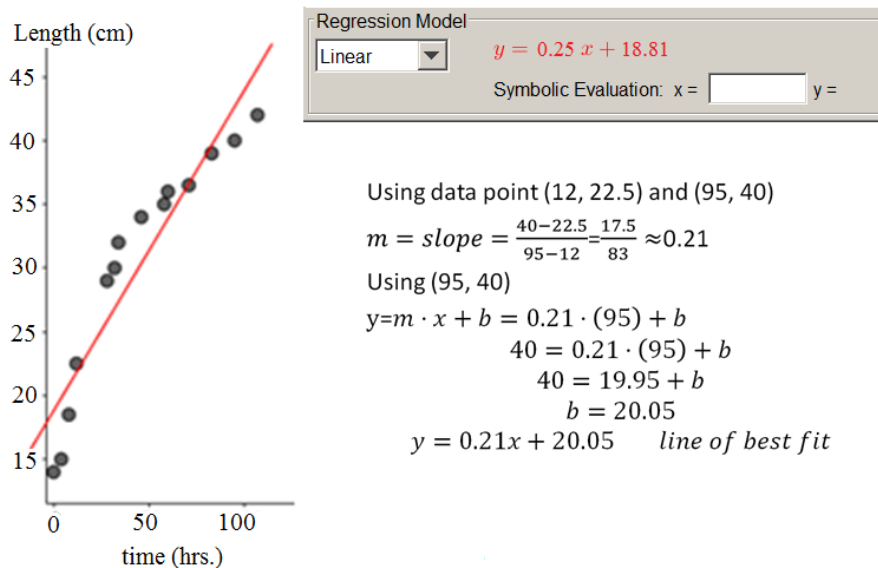$$y = 0.21x + 20.05 \quad line\ of\ best\ fit$$

Fig. 3. Line of best fit, GeoGebra, calculations of line of best fit

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Where β_0 and β_1 are unknown parameters and ε is the random error with a mean 0

and variance σ^2. So, $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X$

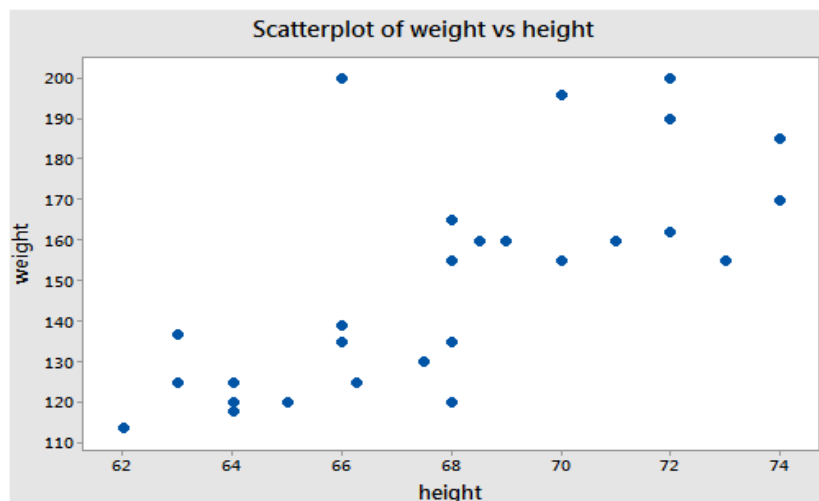Model: $\hat{Y} = -222 + 5.49 \cdot X$

Fig. 4. Model that generates data and scatterplot of simulated data

line of best fit shows regarding the relationship amongst these variables in context. They were instructed to run the model repeatedly and draw inferences about the relationship amongst the variables and the line of best fit. Afterwards, they were asked to revise the model, run the model repeatedly to generate data, and discuss criteria guiding informal inference and the goodness of fit of the model.

Analysis of the data began with viewing the videotapes and transcribing the interviews. Constant comparative analysis (Corbin & Strauss, 2007) was used to analyse a few cases of teachers' responses informed by moving back and forth between viewing the videotapes, and classifying the teachers' responses. Analysis of participants' responses was informed by the teachers' model-based informal inference and the researchers' expectations, which guided the development of the two interviews and generation of the coding scheme. In this chapter, I summarize teachers' conceptions and explanations of the criteria used to plot an informal line of best fit.

4. Results

4.1 Conceptions of Models and Models' Fit

The teachers' responses from the first interview showed the following predominant personal conceptions regarding the informal line of best fit. In the real data branch of the bidirectional modelling diagram of model-based informal inference, all teachers looked at the plotted data on the scatter graph and attempted to build the informal line of best fit. When the teachers engaged with task one, about one third of the teachers (N=6) understood the regression line of best fit as the model that shows the relationship amongst the variables in the population. The largest group of teachers (N=7) viewed the line of best fit as the best representation of all the sample data, where the line represents the data displayed in the scatter plot rather than showing a more general relationship. The remaining teachers (N=4) cited factors contributing to characterise the line of best fit as "typical" such as the line that averages all the positions of all the data points. The term "typical" is one of the measures of centre, thus sometimes was named as average, or median. In the context of inferring about the middle or average of all the data, these teachers seemed to anticipate the line of best fit to be the bivariate equivalent for determining the typical value situated in the middle of the bivariate data, "in such a way that values higher than the average are countered by values lower than the average" according to one eloquently articulated excerpt.

In the real data branch of the bidirectional modelling diagram of model-based informal inference, almost all the teachers' (N=15) responses to the second task, indicated that the regression linear model displayed in a scatter graph, is easy to visualise and the signal is meaningful for data because it eliminates the inherent noise. The remaining teachers (N=2) focused their attention on the feature of the line of best fit as a predictor that enables prediction for data that are not in the data set. To make sense of the situation, one teacher said, "I guess the addition of more bivariate data values will introduce variation that will change the line (regression model)". Nearly all teachers (N=16) believed that the linear model would be modified appropriately, coordinating the structure of the model to the added bivariate data values.

4.2 Criteria for Placing the Informal Line of Best Fit

About one third of the teachers (N= 6) judged linear regression model fit by just measuring the distance from the line to the data points and where the line was closest to the most data points. These teachers explicitly stated that the best location of the line of best fit is when the line is closest to all the data points minimising the distance from points to the line of best fit. Four of the teachers explicitly mentioned an equal number of points above and below the line as a criterion of model fit. However, when the teachers attempted to adjust the line of best fit, most of the examples had roughly (and not exactly) equal numbers of data points above and below the line of best fit. Only two of the teachers further judged the linear regression line as an adequate description of the process in light of the pairing of the data points to determine the line of best fit. These teachers judged that when each point above the line is paired with a point below the line so both points in the pair are equally distant from the line. The teachers expressed this criterion for model fit, especially when they viewed the data points of the height and weight of humans in task 2. Teachers' articulations also give rise to another criterion, the sum deviation criterion. In their original utterances, five teachers very well-articulated that deviations for points above the line and below the line, must sum equally.

4.3 Placement of Informal Line of Best Fit and the Relationship Amongst the Variables in Context

While the teachers in the "model" branch observed the generation of the data and the simulated lines of best fit, the different simulated lines of best fit were notably similar after the model ran repeatedly. Nearly all the teachers (N=16) demonstrated an aggregate view of the data, while the remaining teacher (N=1) seemed particularly attentive to specific points on the line of best fit or points in the data set in order to determine the line. Whereas teachers' inferences about the second model focused predominantly on the wider variation of data, two criteria were predominantly utilised throughout this task: Most teachers (N=14) judged linear regression model fit by just measuring the distance from the line to the data points, and the sum deviation criterion where the line was closest to the most data points.

5. Conclusions

The bidirectional modelling process allows modellers to coordinate the bidirectional modelling process with their conceptions about the informal line of best fit and the criteria they developed to fit informally the line of best fit. The accessibility of the process provides opportunities to conduct an analysis of the inherent variability of the data when modelling with a line, using conceptions, such as the line of best fit represent the data points, or it is typical, or signal, or a predictor or a model. The meanings attributed to the informal line of best fit support understanding of the modelling inherent in formal inferential processes. These meanings that appeared to be situationally rooted in the abstract theory of statistical inference provide further insight into teachers' understanding of a line of best fit, and how they will teach model-based informal inference. The teachers in this study used four criteria (equal number, closest, pairs, and sum deviation) for determining placement of the line of best fit. These criteria must be coordinated with the meanings developed by the teachers when attempted to informally place the line of best fit. This research represents a relatively early iteration of a design cycle of research on teachers' thinking regarding informal inferences. Future research will investigate: a) students' model-based informal inferences about the line of best fit, b) teachers' knowledge about model-based informal reasoning regarding non liner models, and c) students' model-based informal inferences about non-linear models. I anticipate that future research would help with further re-design of instruction that may support students to develop a continuum of knowledge about model-based informal inference and coordinate relationships amongst models of bidirectional modelling processes and model-based informal inference related to these models.

References

Blomhøj, M., & Jensen, T.H. (2007). What's all the fuss about competencies? In W. Blum, P. Galbraith, H. Hen, & M. Niss (Eds.), *Modelling and applications in mathematics education*, 45-56. NY: Springer. https://doi.org/10.1007/978-0-387-29822-1_3

Blum, W., & Leiß, D. (2007). How do students' and teachers deal with modelling problems? In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical modelling: Education, engineering and economics*, 222-231. Chichester: Horwood. https://doi.org/10.1533/9780857099419.5.221

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9-13. https://doi.org/10.3102/0013189X032001009

Cochran, W. G. (1977). *Sampling Techniques*, 3rd Edition. New York: W.W. Norton.

Corbin, J., & Strauss, A. (2007). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Thousand Oaks, CA: Sage.

Engel, J., & Kuntze, S. (2011). From data to functions: Connecting modelling competencies and statistical literacy. In G. Kaiser, R. W. Blum, R. Borromeo Ferri, & G. Stillman (Eds.), *Trends in the teaching and learning of mathematical modelling*, 397–406. New York: Springer. https://doi.org/10.3102/0013189X032001009

English, L. D. (2012). Data modelling with first-grade students. *Educational Studies in Mathematics, 81*(1), 15-30.

Galbraith, P., Stillman, G., & Brown, J. (2010). Turning ideas into modeling problems. In R. Lesh, P.L. Galbraith, C. R. Haines & A. Hurford (Eds.), *Modeling students' mathematical modeling competencies*, 133-1443. New York: Springer. https://doi.org/10.1007/s10649-011-9377-3

TinkerPlots: *Dynamic data exploration (Version 2.0). Emeryville: CA: Key Curriculum Press.*

Konold, C., & Pollatsek, A. (2002). Data analysis as the research of signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259-289. *https://doi.org/10.2307/749741*

Lehrer, R., Kim, M. J., Jones, R. S. (2011). Developing conceptions of statistics by designing measures of distribution. ZDM, *43*(5), 723-736. https://doi.org/10.1007/s11858-011-0347-0

Lehrer, R., & Schauble, L. (2010). What king of explanation is a model? In M. K. Stein (Ed.), *Instructional Explanations in the Disciplines*, 9-22. New York: Springer. https://doi.org/10.1007/978-1-4419-0594-9_2

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The annals of Applied Statistics, 7*(1), 295-318. https://doi.org/10.1007/978-1-4419-0594-9_2

Lu, J. (2016). On randomization-based and regression-based inferences for 2^k factorial designs. *Statistics and Probability Letters, 7*(1), 119, 72-78. https://doi.org/10.1016/j.spl.2016.01.010

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), On the shoulders of giants: A new approach to numeracy 95-137. Washington, DC: National Academy of Sciences.

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: a classroom study.

Mathematical Learning and Thinking, 13(1&2), 27-46. https://doi.org/10.1080/10986065.2011.538302

Pratt, D., & Ainley, J. (2008). Introducing the special issue of informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 3-4.

Prodromou, T. (2012). Connecting experimental probability and theoretical probability. *International Journal on Mathematics Education (ZDM), Springer, 44*(7), 855-868. https://doi.org/10.1007/s11858-012-0469-z

Prodromou, T., & Pratt, D. (2006). The role of causality in the Co-ordination of the two perspectives on distribution within a virtual simulation. *Statistics Education Research Journal, 5*(2), 69-88.

Quetelet, A. (1835). Sur l' Homme et le développement de ses facultés, ou Essai de physique sociale. Paris: Bachelier.

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

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry (with discussion). *International Statistical Review, 67*, 223-265. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Copyrights