A triplet of data/context, mathematical model, and statistical model: Conceptualising data-driven modelling in school mathematics

Takashi Kawakami[1]

[1]Utsunomiya University, Japan; t-kawakami@cc.utsunomiya-u.ac.jp

This study constructs and illustrates a framework for describing and analysing data-driven modelling in school mathematics from the perspectives of mathematical and statistical models. Based on the constructivist view of models and the perception that they are the subject's deterministic or stochastic interpretation, the framework consists of the triplet of data/context, mathematical model, and statistical model, underpinning predictions. An existing research case is used to support the notion that the framework can serve as a tool for describing and analysing aspects of the transition and integration between mathematical and statistical modelling. In addition to being a research tool for linking mathematical modelling research and statistics education research, the framework can be a pedagogical tool for designing activities and teacher intervention in data-driven modelling.

Keywords: Data-driven modelling, mathematical model & modelling, statistical model & modelling

Introduction—Rationale for focusing on both mathematical and statistical models

Data availability is increasing in today's society plagued with elements of uncertainty such as the COVID-19 pandemic, climate change, and natural disasters. Mathematical and statistical models are ? also used in the media for data-driven prediction and decision-making. From the perspective of school mathematics, including statistics education, it is important to develop a comprehensive competency to flexibly use both deterministic reasoning based on mathematical models and non-deterministic/stochastic reasoning based on statistical models to consider data from multiple perspectives and make better predictions and decisions under uncertainty (OECD, 2018). One strategy for fostering such competency is to emphasise or position within school mathematics the learning of modelling that generates, validates, and revises mathematical and statistical models for data-driven prediction. This study calls such modelling *data-driven modelling* (DDM).

Maths and statistics education research discuss *empirical modelling* (EM; Berry & Houston, 1995) as a type of mathematical modelling, and *data modelling* (DM; Lehrer & English, 2018) as a type of statistical modelling, respectively. On the one hand, EM focuses on collecting data through experiments and measurements, making sense of the data, creating function graphs and regression equations that fit the data, and making predictions (Berry & Houston, 1995; Stillman & Brown, 2021). Mathematical models, such as functions, are generated through EM. Hence, the focus in EM is typically on tensions between data and mathematical models. On the other hand, DM focuses on data-centric activities, creating empirical distribution models from real-world situations with variability and predicting distributions (Lehrer & English, 2018). Statistical models, such as statistical graphs and statistics that represent variation, are generated through DM. Hence, DM typically focuses on tensions between data and statistical models. Thus, mathematical and statistical modelling research, by nature, have either focused on tensions between data and mathematical models or on tensions between data and statistical models.

Empirical M : EM: collecting / making sense of / create regression functions to predict..
Data M : DM: data centric activities / empirical distribution models / predict distributions

By contrast, focusing on the interaction between mathematical and statistical models in DDM may offer at least two advantages in research and practice. First, it can conceptualise and capture the dynamic aspects of transitions between deterministic and non-deterministic/stochastic reasoning using data as a starting point or mediator. Despite recent modelling research focusing on both mathematics and statistics (e.g. English & Watson, 2018; Kawakami & Mineno, 2021), learners' models have been lumped together under the single term 'model' and such a dynamic aspect has not been fully clarified. Stochastic reasoning is essential for examining the data and considering the imperfections in one's thinking (OECD, 2018); however, learners tend to be biased towards deterministic reasoning (e.g. Casey & Nagle, 2016). Second, it can generate perspectives for more precisely capturing learners' DDM (which integrates EM and DM characteristics) and designing teaching materials and teachers' interventions of DDM, contributing to theorising the relationship between mathematical and statistical modelling (e.g. Ärlebäck & Kawakami, 2023) and linking maths and statistics education research (e.g. Makar & Rubin, 2018).

Statistisches Modellieren kommt momentan zu kurz
⇒ DDM ermöglicht flexible Herangehensweise um die Beziehung zw. math. & stochast. Modellieren heraus zu arbeiten

This theoretical study aims to construct and illustrate a framework for describing and analysing DDM in school mathematics from the perspectives of mathematical and statistical models. It is an extended version of Kawakami and Saeki (2022a), including another illustrative example of the framework and elaborating the framework discussion.

Data-driven modelling with mathematical and statistical models in view

Data
1 – nah am Context
2 – ermöglicht Argumentation und Entscheidungen
3 – besteht aus Signal und Rauschen

Data/context, mathematical model, and statistical model concepts

Data refer to a sequence of values of a variable or a sequence of pairs of values of multiple variables. The concept has three characteristics. First, data are closely related to real-world context (Cobb & Moore, 1997), which requires understanding and interpreting data based in the real-world context. This study uses the term *data/context* when emphasising the closeness of data and real-world context. Second, data provide evidence for reasoning and decision making (Page, 2018), which means that data is both a source of representations (i.e. models) for reasoning and decision-making and a comparison target for validating the validity of the representations (i.e. models). Third, the structure of data consists of *signal* and *noise* (Konold & Pollatsek, 2002). Signal refers to regular patterns or central tendencies in the data, whereas noise refers to irregular random patterns or errors in the data.

A *model* refers to a representation of structure in a given system (Hestenes, 2010) and reflects the subject's series of interpretations of the object (Lesh & Doerr, 2003). In this constructivist view of models, structure is not seen as something that exists in a fixed way, but as something that changes and is constituted in the subject (Piaget, 1968). Based on the constructivist view of models and the way of perceiving mathematical or statistical models as the subject's deterministic or stochastic interpretation (e.g. Dvir & Ben-Zvi, 2023; Kondo, 1976), this study did not take the same representation absolutely, but *relatively* as a mathematical or statistical model, depending on the learner's intention and interpretation of the model. By *mathematical model*, this study means a representation of the signal in the data, reflecting the subject's deterministic interpretation of the data and context. This model is generated by viewing the behaviour observed in the data/context as a certain outcome or pattern that results from a certain cause, such as a linear model *y*=*ax*+*b* (*a* and *b*

holistischer Ansatz. Modelle in den Köpfen der Menschen sind nicht fix

Modell: Eine Repräsentation einer Struktur, welche die Interpretation des Modellierenden wiedergibt
Mathematisches Modell: Repräsentation des Signals in den Daten, welche die Interpretation und Einbindung in den Kontext durch den Modellierenden wiedergibt.
Statistisches Modell: Repräsentation des Rauschens in den Daten, welche die nicht-deterministische/Stochastische Interpretation der Daten und des Kontextes des Modellierenden wiedergibt.

are parameters), where the value of variable y is determined if the value of variable x is determined. Typical examples of mathematical models that can be generated through DDM activities in school mathematics include graphical and algebraic representations of functions (e.g. proportional, inverse proportional, linear, quadratic). By *statistical model,* this study refers to a representation of the noise in the data, reflecting the subject's non-deterministic/stochastic interpretation of data and the context. This model is generated by viewing the behaviour observed in the data/context as a variable outcome or pattern, even from the same cause, such as a linear model $y=ax+b+\varepsilon$ (a and b are parameters) where the value of variable y is not determined even if the value of variable x is determined and is distributed by a random error ε. Typical examples of statistical models that can be generated through DDM activities in school mathematics include statistical graphs (e.g. dot plots, histograms, and scatter plots), statistics (e.g. representative values and standard deviations), and statistical probabilities. Mathematical and statistical models are also called deterministic and probabilistic models, respectively.

DDM concept

This study specifies the DDM concept that focuses on both mathematical and statistical models. In doing so, this study draws on four components of modelling with data, as identified by English and Watson (2018), with the intention of linking mathematics and statistics education. The first is *boundary interactions between mathematics and statistics*, which is the generalisation of mathematical and statistical models not only to pursue the meaning of specific data/contexts, but also to apply them to new data/contexts. The second is *interpreting and reinterpreting problem contexts and questions,* which is the iterative interpretation and formulation of complex real-world contexts and unstructured questions. The third is *interpretating, organising and operating on data in model construction*. This is the generation, validation, and revision of models that describe and explain trends and variability in the data through the combination of mathematical and statistical reasoning applied to the data. Although English and Watson (2018) did not distinguish between mathematical and statistical models, the constructed models are expected to include both mathematical and statistical models (see the first component). The fourth is *drawing informal inferences,* which is making predictions about unknown data from the data at hand (Makar & Rubin, 2018) as the goal of the modelling activity, without assuming formal statistical inference based on probability distributions. Based on those four components, this study defines DDM as a series of activities of generating, validating, and revising models (i.e. mathematical and statistical models) that describe and explain trends and variability in data by repeatedly interpreting a real-world context and interpreting, organising, and processing data to make better predictions (i.e. informal inferences) for the real world.

DDM characteristics

This study summarises DDM characteristics, contrasting them to those of EM and DM. In EM, the data set is the object of modelling; however, it is also important to consider the real-world context (Stillman & Brown, 2021). EM is generally application-oriented and is viewed as the practical orientation (Berry & Houston, 1995) as it solves real-world problems with prediction. Meanwhile, DM allows for integrated activities that interrelate application- and structure-oriented understanding and developing concepts for better prediction (Lehrer & English, 2018).

EM and DM share similarities with DDM in that they emphasise real-world context, focus on the relationship between data/context and models, and set up the purpose of data-based prediction. At the same time, EM and DM differ from DDM, which focuses on both mathematical and statistical models, in that they focus on either mathematical or statistical models. DDM focuses on tensions between data and models, especially both mathematical and statistical models. Both DM and DDM allow for the integration of application- and structure-orientation. In prediction, they extend the range of applicability of the model from the sample at hand to a sample distribution or population not at hand, as well as to generalise the model. Table 1 summarises the abovementioned characteristics of EM, DM, and DDM, and shows that DDM integrates both EM and DM characteristics.

Table 1: Summaries of EM, DM, and DDM characteristics

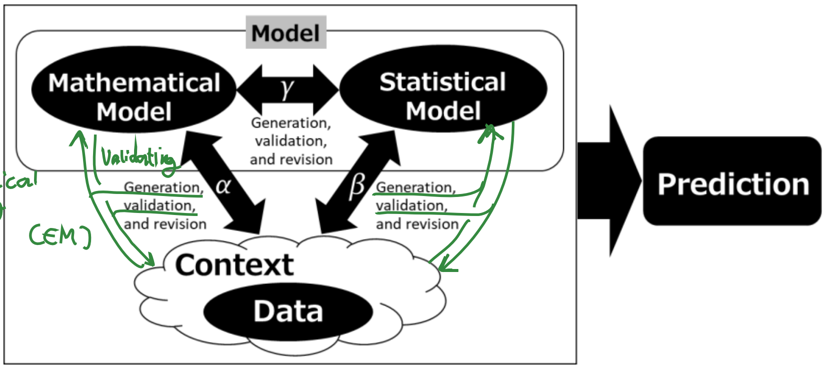
Characteristics	**EM (e.g. Berry & Houston, 1995)**	**DM (Lehrer & English, 2018)**	**DDM**
Emphasis on real-world context	✓	✓	✓
Aims for data-based prediction	✓	✓	✓
Focus on tensions between data and mathematical models	✓		✓
Focus on tensions between data and statistical models		✓	✓
Application-oriented (i.e. real-world problem solving with predictions)	✓	✓	✓
Structure-oriented (i.e. concept development for better predictions)		✓	✓

Proposed framework for describing and analysing DDM

This study proposes a framework (Figure 1) to describe and analyse DDM in school mathematics, consisting of three transitions between: (α) data/context and mathematical models, (β) data/context and statistical models, and (γ) mathematical and statistical models (Kawakami & Saeki, 2022a).

Mathematical modelling

(EM)


Model
Mathematical Model
Statistical Model
γ
Generation, validation, and revision
Validating
Generation, validation, and revision
α
β
Generation, validation, and revision
Context
Data
Prediction

Mathematical Model <=> Context

Figure 1: A framework for describing and analysing DDM

Transition α refers to the activity of moving back and forth between the data/context and the mathematical model. The transition from data/context to a mathematical model corresponds to the generation of a mathematical model representing the signal based on the data/context. The transition from the mathematical model to the data/context involves validating the generated mathematical model based on the data/context, and, if necessary, revising the mathematical model. EM is an activity focused on transition α. *Transition β* refers to activities that move back and forth between the data/context and the statistical model. Moving from data/context to a mathematical model involves generating a statistical model that represents the noise based on the data/context. The transition from the statistical model to the data/context involves validating the generated statistical model based on the data/context and, if necessary, revising the model. DM is an activity focused on transition β. *Transition γ* refers to the activity of moving back and forth between mathematical and statistical models. Referents of the models may or may not remain unchanged. On the one hand, generation of a statistical model based on a mathematical model (mathematical model → statistical model) involves the stochastic representation and interpretation of a mathematical model, or the generation of a new statistical model based on the representation and interpretation. On the other hand, generation of a mathematical model based on a statistical model (mathematical model ← statistical model) corresponds to the deterministic representation and interpretation of a statistical model, or the generation of a new mathematical model by the representation and interpretation. In this way, transition γ involves generating models from models rather than data/context. In doing so, model objectification (e.g. Lesh & Doerr, 2003) is performed to reflect on the generated models and make the models the objects of thought to transform them into more general and powerful models.

Reihenfolge ist nicht festgelegt

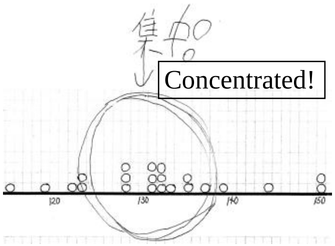
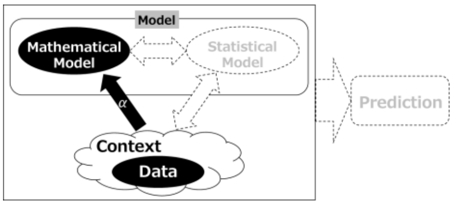
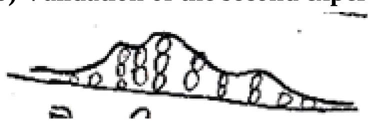
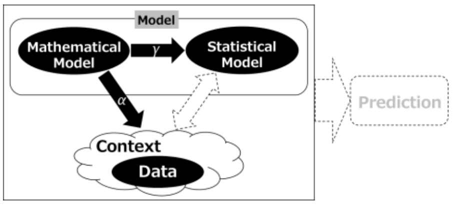
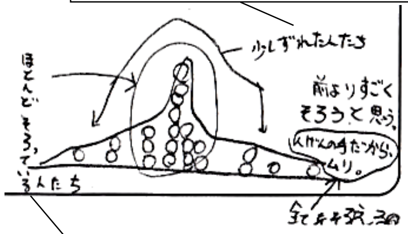
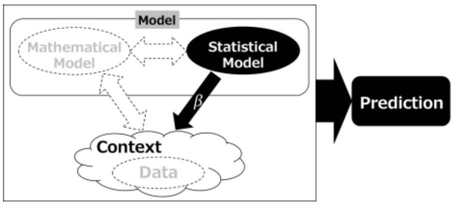
Note that no sequentiality is assumed for transitions α, β, and γ. When generating a model from data/context, it is conceivable that transition α or β is followed by transition γ. It is possible to start with transition γ if DDM is started by analysing an existing model. Transition γ may be performed together with transitions α and β. Hence, the validation and revision of statistical and mathematical models can occur not only in transition γ but also in transitions α and β.

Illustrative example of the framework

nur ein Schüler

Table 2 illustrates the framework with the case of student Ayu's prediction in Kawakami's (2017) lessons. In the lessons, 10–11-year-old students engaged in 'paper helicopter experimentation' that included predicting the flight time distribution of a paper helicopter dropped from a certain height and validating and revising the prediction through measuring the flight time. Ayu's initial prediction was limited to a focus on the distribution components (i.e. centre and density), but only in the final prediction did she explicitly associate the components (i.e. centre, density, and shape) with each other and the context. Kawakami (2017) focused on the transformation of learners' models of data distribution using the single model concept. The current study identifies mathematical and statistical models based on Ayu's worksheet description and the framework. Table 2 shows that the generation of the mathematical model contributed to the generation of the statistical model, leading to predictions that considered signal, noise, and context. Transition γ between mathematical and statistical models led to a greater choice of models for learners to use in making predictions and the promotion of non-deterministic/stochastic reasoning based on deterministic reasoning. Moreover, Table 2(b) shows that as the second experiment results were being validated and the validity of the initial prediction was being examined, transition γ was caused by the objectification of the dot plot illustrating the results. Mathematical modelling research has shown that learners deepen their understanding of models created in the mathematical world and extend the applicability of models to similar and new situations (Lesh & Doerr, 2003). Statistical modelling research has also illustrated DM extending the

Table 2: Illustration of the framework using the case of Ayu’s prediction in Kawakami (2017)

Ayu’s worksheet descriptions	*Explanations using the framework*
(a) Second experiment and visualisation of results The paper helicopter  Dot plot of flight-time data Ayu conducted the second paper helicopter experiment and organised the results into a mean and dot plot of flight time. In the dot plot, data were converted into 0.01 bps. On the dot plot, she enclosed the intervals where the data were concentrated.	**(a) Generation of a mathematical model based on data/context [transition *α*]**  Model Mathematical Model Statistical Model α Prediction Context Data Here, the dot plot, along with the mean, was interpreted deterministically with a focus on organising the data at hand. The dot plot also represented the signal (the interval in which the data were concentrated, i.e. the central tendency). This dot plot can be considered a mathematical model. Thus, based on the data collected through the experiment, a mathematical model was generated.
(b) Validation of the second experiment results  Idealised dot plot of the second experiment results ‘Although I thought the dots would clump more, the dots clumped only a little. I think the timing error was large in the real distribution.’ Validation of the results and her initial prediction In validating the results of the second experiment, Ayu compared the results of her first experiment with the results of the second and idealised the distribution using the shape concept, enclosing the distribution even where no dots were drawn. She also considered the difference between her initial prediction and real distribution and attributed the difference to the timing error.	**(b) Generation of a statistical model based on a mathematical model [transition *γ*] and validation of a mathematical model based on data/context [transition *α*]**  Model Mathematical Model γ Statistical Model α Prediction Context Data Here, the idealised dot plot was interpreted stochastically to mean that areas without dots might exist if the sample size were increased. It represented noise (the measurement error) and can be considered a statistical model. By surrounding the approximate shape of the distribution, the dot plot in (a) (as a mathematical model that represented the data at hand) was interpreted stochastically to generate a statistical model that also represented data not at hand. The mathematical model was also validated based on real data and measurement context.
(c) The final predication of distribution Cases of the slight timing lag of measurement  Cases of the exact timing of measurement Ayu predicted that a more aligned stopwatch timing would result in a nearly symmetric unimodal distribution. She surrounded the shape of the dot plot, making sense of the centre, left, and right hem of the distribution in terms of the measurement error.	**(c) Prediction based on a statistical model associated with contextual sense making [transition *β*]**  Model Mathematical Model Statistical Model β Prediction Context Data Here, the dot plot was interpreted stochastically and consideration was given to the shape of the population variability. It represented noise (the measurement error) as well as signal (symmetric unimodal shape; the central tendency) and was associated with measurement context (variation source). The dot plot here can be considered a statistical model. The statistical model was used for the final prediction.

applicability of models from the sample at hand to sample distributions and populations not at hand (Lehrer & English, 2018). As shown in Table 2, a learner's DDM activities can include the same applicability. Ayu's DDM activities started from the data and transitioned between mathematical and statistical modelling (i.e. EM and DM) or integrated the two. Thus, by analysing the case of the paper helicopter with, in general, undifferentiated modelling of data-driven situations using the framework, we can describe in detail the diverse and multifaceted thinking processes and thinking outcomes about data of learners who use deterministic and non-deterministic/stochastic reasoning flexibly.

Conclusion—Feasibility of using the framework for research and practice

While this study developed the framework using only one limited case study, Kawakami and Saeki (2022a) illustrated the framework using another case study of four groups of 12–13-year-old students. In summary, I discuss the feasibility of using the framework for both research and practice. In terms of research implications, the framework suggests the comprehensive competency required of learners in modelling data-rich situations, that is, the competency to move adaptively between data/context, mathematical, and statistical models for better predictions and decision making. The competency to engage in DDM is crucial for citizens living in a big data and AI world (OECD, 2018). Additionally, we can reconceptualise undifferentiated modelling activities in data-rich situations in school mathematics with the perspectives of mathematical and statistical models. The three main types of such modelling activities are: EM, which focuses on tensions between data and mathematical models; DM, which focuses on tensions between data and statistical models; and DDM, which focuses on tensions among data, mathematical models, and statistical models. This study's characterisations provide a holistic lens through which to relativise the characteristics of existing research on modelling in data-rich situations. Since DDM includes the characteristics of EM and DM, the framework can be a research tool for linking mathematical modelling research and statistics education research. As practical implications, the framework implies basic DDM activity types in terms of mathematical and statistical models and application- and structure-orientation. The example in this study corresponds to a statistical-/structure-oriented activity whose focus is to develop understanding of mathematical and/or statistical concepts with added emphasis on predictions using statistical models generated from mathematical models. The framework also suggests opportunities and methods of teacher intervention in DDM in terms of transitions α, β, and γ. This study suggests model objectification (e.g. Lesh & Doerr, 2003) as an intervention for transition γ. DDM includes a diverse range of activities, including societal decision-oriented and interdisciplinary activities (Kawakami & Saeki, 2022b, forthcoming). This framework requires further elaboration to demonstrate the possibilities of different DDM activities in line with a data-rich world.

Acknowledgements

I thank my primary PhD supervisor Professor Akihiko Saeki, and my secondary supervisor Professor Kazuhiko Nunokawa for their useful suggestions during this research. This study was supported by JSPS KAKENHI Grant Numbers JP21K02513.

References

Ärlebäck, J. B., & Kawakami, T. (2023). The relationship between statistics, statistical modelling and mathematical modelling. In G. Greefrath, S. Carreira, & G. Stillman (Eds.), *Advancing and* *consolidating mathematical modelling: Research from ICME-14* (pp. 293–309). Springer. https://doi.org/j6hr

Berry, J., & Houston, K. (1995). *Mathematical modelling*. Edward Arnold.

Casey, S. A., & Nagle, C. (2016). Students' use of slope conceptualizations when reasoning about the line of best fit. *Educational Studies in Mathematics*, *92*(2), 163–177. https://doi.org/jrjh

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801–823. https://doi.org/10.2307/2975286

Dvir, M., & Ben-Zvi, D. (2023). Informal statistical models and modeling. *Mathematical Thinking and Learning*, *25*(1), 79–99. https://doi.org/jrjj

English, L., & Watson, J. (2018). Modelling with authentic data in sixth grade. *ZDM Mathematics Education*, *50*(1-2), 103–115. https://doi.org/gpb58r

Hestenes, D. (2010). Modeling theory for math and science education. In R. Lesh, P. Galbraith, C. Haines, & A. Hurford, (Eds.), *Modeling students' mathematical modeling competencies* (pp. 13–41). Springer. https://doi.org/dtpqtr

Kawakami, T. (2017). Combining models related to data distribution through productive experimentation. In G. A. Stillman, W. Blum, & G. Kaiser (Eds.), *Mathematical modelling and applications* (pp. 95–105). Springer. https://doi.org/hnw2

Kawakami, T., & Mineno, K. (2021). Data-based modelling to combine mathematical, statistical, and contextual approaches: Focusing on ninth-grade students. In F. K. S. Leung et al. (Eds.), *Mathematical modelling education in east and west* (pp. 389–400). Springer. https://doi.org/jrp5

Kawakami, T., & Saeki, A. (2022a). A framework for describing and analysing data-driven modelling activities in school mathematics: From the perspectives of mathematical and statistical models. *Journal of Science Education in Japan*, *46*(4), 421–437. (in Japanese) https://doi.org/jtn3

Kawakami, T., & Saeki, A. (2022b). *The role of mathematical and statistical models in data-driven modelling: A prescriptive modelling perspective*. Presentation at the 20th International Conference on the Teaching of Mathematical Modelling, 24–27 September 2022, Online. https://doi.org/jr5h

Kawakami, T., & Saeki, A. (forthcoming). Extending data-driven modelling from school mathematics to school STEM education. In J. Anderson, & K. Makar (Eds.), *The contribution of mathematics to school STEM education: Current understanding*. Springer.

Kondo, J. (1976). *Sugaku moderu: Gensho no sushikika* [Mathematical models: Mathematization of phenomena]. Maruzen. (in Japanese)

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

Lehrer, R., & English, L. (2018). Introducing children to modeling variability. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 229–259). Springer. https://doi.org/jr5j

Lesh, R. A., & Doerr, H. M. (Eds.). (2003). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Routledge. https://doi.org/jr5k

Makar, K., & Rubin, A. (2018). Learning about statistical inference. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 261–294). Springer. https://doi.org/jr5m

OECD (2018). *PISA2022 mathematics framework (Draft)*. https://pisa2022-maths.oecd.org

Page, S. (2018). *The model thinker: What you need to know to make data work for you*. Basic Books.

Piaget, J. (1968). *Structuralism*. Psychology Press. https://doi.org/jr5n

Stillman, G., & Brown, J. (2021). Modeling the phenomenon versus modeling the data set. *Mathematical Thinking and Learning*. https://doi.org/jr5p