Roles of Mathematical and Statistical Models in Data-Driven Predictions in an Integrated STEM Context

Takashi Kawakami
Utsunomiya University, Japan
t-kawakami@cc.utsunomiya-u.ac.jp

Akihiko Saeki
Kanazawa Institute of Technology, Japan
asaeki@neptune.kanazawa-it.ac.jp

This study elaborates on the pivotal roles of mathematical and statistical models in data-driven predictions in an integrated STEM context using the case of Year 4 students: (i) *a descriptive means* to describe the features of trends and variability of data and (ii) *an explanatory means* to explain causal relationships behind data. These roles are linked to models in other STEM subjects (i.e., prototypes and scientific models) and the application and development of STEM content knowledge. The results contribute to a better understanding of the role of mathematics/statistics in STEM education.

Predictions are found everywhere in life, society, and science. During the global COVID-19 pandemic, daily data on the number of positive cases, severe cases, and deaths were published in graphs and tables, and it became routine to keep track of the current infection situation and predict future waves of infection. *Data-driven prediction* by using data, mathematics, statistics, and interdisciplinary knowledge to predict and validate complex and uncertain events is indispensable for today's citizens and societies (e.g., Geiger et al., 2023). To provide a vehicle and platform for such data-driven predictions, modelling processes involving the generation, evaluation, and revision of *mathematical models* (deterministic representations) and *statistical models* (non-deterministic/stochastic representations) in data-rich interdisciplinary contexts are gaining attention in Science, Technology, Engineering and Mathematics (STEM) education from a mathematics education perspective (e.g., English, 2023). However, the literature does not clearly explain how and to what extent students use mathematical and statistical models for data-driven predictions in an integrated STEM context.

Unravelling the pivotal roles of mathematical and statistical models in STEM education offers two advantages for research and practice. First, it could elaborate on and advance the role of mathematics/statistics in integrated STEM education (English, 2016) as well as in mathematics curricula, such as ACARA (2022) and MEXT (2018), which emphasise mathematical modelling, statistical investigation, and STEM education. Second, it has implications on developing students' epistemic knowledge about the nature and role of models and representations in STEM disciplines, which are essential for STEM competencies (Tytler, 2020). Therefore, this study elaborates on the roles of mathematical and statistical models in data-driven predictions in an integrated STEM context using the case of Year 4 students.

Conceptual Framework

Data-Driven Prediction

The potential for introducing *data-driven prediction* from the primary school years has been identified by mathematics and statistics education research. *Informal statistical inference* (ISI), in which trends and variations in unknown data are predicted and generalised without adopting formal statistical procedures and methods, is actively studied since primary school years (Makar & Rubin, 2018). For instance, Oslington et al. (2023) highlighted primary school students' predictive reasoning as part of the ISI with representations of patterns such as seasonal trends and variability in data to predict temperature using tables, line graphs, and bar graphs.

Moreover, data-driven predictions are important in an integrated STEM context. Watson et al. (2023) conducted statistical investigations using ISI and technology with primary school

(2024). In J. Višňovská, E. Ross, & S. Getenet (Eds.), *Surfing the waves of mathematics education. Proceedings of the 46th annual conference of the Mathematics Education Research Group of Australasia* (pp. 311–318). Gold Coast: MERGA.

students in a STEM context, revealing their representations, predictions, and understandings of variation. English (2023) implemented modelling involving data-driven predictions with mathematics and statistics for primary school students in an integrated STEM context. It explored how they applied multidisciplinary knowledge of mathematics, statistics, and science through predictions. Aridor et al. (2023) proposed a framework to describe the interactions between statistical reasoning, scientific reasoning, and the nature of scientific understanding using the case of citizen science.

Research (Aridor et al., 2023; English, 2023; Oslington et al., 2023) suggests that it is essential to use *deterministic* reasoning flexibly, based on mathematical models and *non-deterministic/stochastic* reasoning, through statistical model use, to consider data from multiple perspectives and make more reliable predictions for better decision-making. Non-deterministic/stochastic reasoning raises awareness of the limitations of human decision-making and provides an opportunity for critical reflection. Conversely, deterministic reasoning is required when predicting maximum certainty or controlling for uncertainty by seeking conditions that reduce variability. However, few research has demonstrated the roles of mathematical and statistical models in data-driven predictions in an integrated STEM context.

Interdisciplinary Data-Driven Modelling and Functions of Mathematical and Statistical Models

We adopted *interdisciplinary data-driven modelling* (IDDM) considering mathematical and statistical models in data-driven predictions in an integrated STEM context (Kawakami, 2023a, 2023b; Kawakami & Saeki, in press). The IDDM generates, validates, and revises mathematical and statistical models and models in other STEM subjects (science, technology, and engineering) based on data/context to make better predictions (Kawakami & Saeki, in press). Data have a structure comprising a deterministic aspect (*signal*) focused on exact numbers and causal explanations with certainty and a non-deterministic/stochastic aspect (*noise*) focused on uncertainty and variability (Innabi et al., 2023). A model refers to a representation of the structure of a given system and a reflection of the modeller's series of interpretations of an object (Hestenes, 2010). A *mathematical model* refers to a representation of the signal inherent in the data, reflecting the modeller's deterministic interpretation of the data and context (Kawakami, 2023b). A typical example is the linear model $y = ax+b$ (a and b are parameters), where the value of the variable y can be determined if the value of the variable x is determined. A *statistical model* refers to a representation of the noise inherent in the data, reflecting the modeller's non-deterministic/stochastic interpretation of the data and context (Kawakami, 2023b). A typical example is the linear model $y = ax+b+\varepsilon$ (a and b are parameters), where the value of the variable y cannot be determined even if the value of the variable x is determined and distributed by a random error ε. *Models in other STEM subjects* involve modellers' representations and interpretations of data, which are relevant to big ideas in STEM disciplines (Kawakami & Saeki, in press), such as scientific models (e.g., motion models of a falling body and structural models of seeds) and engineering models (e.g., scale model/prototypes).

Mathematical and statistical models describe phenomena and explain their prediction mechanisms. Ärlebäck and Doerr (2020) showed that models serve as *descriptive means* and *explanatory means*. The former is a function of understanding and describing the behaviour of events. The latter is a function of explaining the structure of events and the mechanisms of their structure in a unified and comprehensive way and elucidating why events behave as they do. They pointed out that a unified explanation of several events using the same model leads to the idea of generalisation, and it is necessary to combine several models with different perspectives to provide a comprehensive explanation of a single event.

Additionally, the descriptive and explanatory functions of mathematical and statistical models are essential in an integrated STEM context. To build a causal narrative about why a

phenomenon occurs, Baptista et al. (2023), who identified the core features of a reasonable explanation for a STEM problem, required students to describe what was happening in a given scenario (by summarising patterns in the data), make connections between observations and mathematical or statistical models, and use scientific ideas. Thus, mathematical and statistical models are expected to contribute to data-driven predictions in an integrated STEM context through IDDM by describing and reading the signal and noise characteristics in the data and explaining them in a unified and comprehensive manner (Table 1).

Table 1

Functions of Mathematical and Statistical Models that Could Contribute to Data-Driven Prediction

Functions	Descriptions
Descriptive means	Mathematical and statistical models, such as graphs and statistics, provide an external representation of the signal and noise characteristics inherent in data, providing an understanding of how the data behaves
Explanatory means	Mathematical and statistical models, such as graphs and statistics, provide a unified and comprehensive explanation of the signal and noise characteristics inherent in data, providing an understanding of the behaviour of data and the mechanisms and causal relationship of events behind the data

Given the theoretical framework of the functions of the mathematical and statistical models in Table 1, we formulated and addressed the following research question:

- Considering the model functions of descriptive and explanatory means, how do students use mathematical and statistical models when making predictions through IDDM?

Research Design

Setting, Participants, and Context

To answer the research question, we used data from the IDDM practice implemented in Year 4, where students used mathematical and statistical models and models in other STEM subjects for data-based predictions. An overview of this practice is given in Kawakami and Saeki (in press); however, this study is substantially different from our previous work in that it analyses the role of models in data-driven predictions in practice.

The participants were students (n = 30) from a Year 4 class (aged 9–10 years) in a public primary school in Japan. They had learned about bar graphs, line graphs, and two-dimensional tables. However, they were unaware of representative values, dot plots, and histograms. The practice comprised nine 45-minute lessons in mathematics and cross-curricular enquiry classes and addressed the *Seed Dispersal Task* (Figure 1), incorporating data-driven predictions into Fitzallen et al.'s (2019) seed dispersal material for integrated STEM education.

Figure 1

Seed Dispersal Task (Partial)

> The buckleya lanceolata seeds (Figure 2a) come in various sizes. The speed at which they fall seems to vary depending on the seeds. Therefore, how can the flight time of the seeds be increased?
>
> **Sub-task:** After several experiments with the seed prototype (Figure 2b) to measure the flight time, predict what will happen to it if the slit length of the prototype is varied, as shown in Figure 2c. After making a prediction, validate it through experiments.

The goal of the task was to redesign the shape of the buckleya lanceolata seed (Figure 2a) to maximise flight time. After several experiments that measured the flight times of seed prototypes that behaved similarly to the seeds when they fell (Figures 2b and 2c), the sub-task involved predicting trends in flight times with the prototype shape as a variable and validating the predictions with real data. In the sub-task, the students generated mathematical and

statistical models (e.g., line graphs) and models in other STEM subjects. For example, the students interpreted data trends deterministically and stochastically in relation to seed prototype size, weight, and shape as engineering models. Students also generated scientific models of the motion of a falling body under air resistance and the structure and function of the seeds.

Figure 2

Buckleya Lanceolata Seed and the Seed Prototypes (Photos Taken by Shohei Chiba)

a)  **b)**  **c)** 

To answer the research question, we focused on a sub-task involving prediction. In the lessons, before working on the sub-task, the students experienced dropping prototypes with vertical lengths of 15 cm and 20 cm and predicted the change in flight time as the vertical length increased, by drawing line graphs on the worksheets. After making predictions, they collected data on the flight times of the prototype with longer vertical lengths, plotted the datasets on a line graph, and compared the predicted graph with the actual data to validate their predictions.

In the sub-task, the students collected data by experimenting with the flight time of a prototype with a fixed vertical length and slit lengths of 3 cm, 6 cm, and 9 cm (Figure 2c). Then, they plotted these data online graphs and presented their predictions regarding the change in flight time when slit length increased. They drew line graphs on the worksheet and calculated the differences in the data (*Prediction*). Once the predictions were made, they collected data on the flight time of the prototype with a further increase in slit length, plotted the data on a line graph, and validated their predictions by comparing the predicted graph they made with the one containing the actual data (*Validation*).

Data Collection and Analysis

We analysed 30 students' worksheet excerpts in the *Prediction* and *Validation* activities and used a post-class interview protocol for complementary analysis. The prediction intention could also be written in the validation statement; therefore, it was included in the analysis. These data were analysed in three coding phases to answer the research question. The first author performed these coding phases, and the second author validated them. The differences in interpretation between the two authors were discussed until an agreement was reached. The analyses were revised as necessary.

In Phase 1, we identified and coded the mathematical and statistical models generated in the sub-task based on the framework of our study. In this analysis, the exact representation was not taken absolutely but relatively as a mathematical or statistical model depending on the student's intention to create and interpret the representation. For example, if a student interpreted a line graph deterministically, the model was taken as mathematical; if interpreted non-deterministically or stochastically, the model was taken as a statistical one.

In Phase 2, we examined whether these models' functions were descriptive or explanatory, based on Table 1, and coded them accordingly. In this analysis, the descriptions of the features of the trends and variability of the flight time data were judged as the emergence of the descriptive function. The use of the models to explain the causal relationship that changes flight time and its consistency with the results of previous experiments was judged as the emergence of the explanatory function.

In Phase 3, we disaggregated the interdisciplinary aspects of students' descriptive or explanatory use of mathematical and statistical models, focusing on the inclusion of models in

other STEM subjects (i.e., engineering models such as prototypes and scientific models such as the air resistance model).

Results

To varying degrees, all participating students used a line graph as either a mathematical or statistical model to predict the flight times of seed prototypes. Considering the functions of the mathematical and statistical models (Table 1), we classified students' use of mathematical and statistical models into three types (Table 2).

Table 2

*Types of Mathematical and Statistical Model Use (*n = *30)*

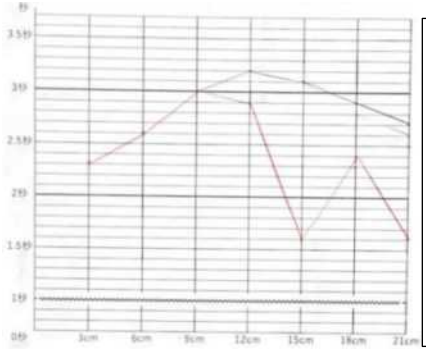
Types	Mathematical model	Statistical model	# (%)
1	Descriptive means	N/A	5 (17%)
2	Descriptive and explanatory means	N/A	20 (66%)
3	N/A	Descriptive and explanatory means	5 (17%)

Type 1 comprised students who used a mathematical model as a descriptive method. Students belonging to this type used the line graph only as a means of reading trends (signal) in the already collected data, but only superficially read the graphs (e.g., "As the flight time was rising, I thought it would continue to rise").

Type 2 comprised students who used a mathematical model for descriptive and explanatory purposes. This was the most common type. Students belonging to this type used line graphs as a means of reading trends (signals) in the already collected data and also as a means of asserting the reasonableness of predictions. They explained the causal relationship between the changing flight time and consistency with experimental results for different horizontal lengths using the different slopes of the graphs. Figure 3 shows an example of the Type 2 model use, demonstrated in a worksheet by Hata (student pseudonym), where on the left is a line graph of the real data and their prediction, and on the right is the reason for the prediction. They described the relationship between feather length and flight time from a line graph of the experimental results before the vertical length was stretched ("In the previous experiment of stretching the vertical length, if the length was stretched too far, the flight time became shorter"). Additionally, they explained why the slope of the flight time graph varied based on the prototype (e.g., "the weight of the prototype remained the same, so I thought the graph would not drop too steeply").

Figure 3

Descriptive and Explanatory Use of a Line Graph as a Mathematical Model From Hata's Worksheet



In the previous experiment of stretching the vertical length, if the length was stretched too far, the flight time became shorter, so I thought that if the slit was made too long, the flight time would become shorter. But the weight of the prototype remained the same, so I thought the graph would not drop too steeply. Moreover, I thought that the lower part of the prototype would become too short and the leaf (wing) part too long, so it wouldn't fly very well. …I thought that the graph would fall gradually, because the part of the prototype that is in contact with the wind would become larger.

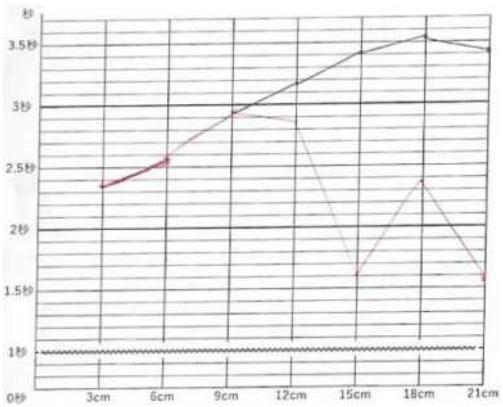
Note. Black graph: Predictive data; Red graph: Real data; Horizontal axis: Prototype slit length; Vertical axis: Flight time.

Type 3 comprised students who used a statistical model as a descriptive and explanatory means. Students belonging to this type used line graphs as a means of reading variability (noise)

in the data already collected and also as a means of asserting the reasonableness of predictions. Figure 4 shows an example of the Type 3 model used, demonstrated in a worksheet by Aki (student pseudonym). They described variability in the flight time from a line graph ("the flight time increased between 0.20 and 0.39 s") and utilised it to make predictions with an awareness of the range of values ("the flight time would also increase with a further increase in slit length, roughly between 0.20 and 0.39 s"). Referring to the geometry of the prototype, they also explained in a unified way the results of previous experiments with different vertical lengths and the results of the current experiments with different slit lengths ("The reason for lowering the graph at the end was the same as in the previous experiment with different vertical lengths: I thought that the longer the slit, the faster it would fall").

Figure 4

Descriptive and Explanatory Use of a Line Graph as a Statistical Model From Aki's Worksheet



Since the flight time increased between 0.20 and 0.39 s, I predicted that the flight time would also increase with a further increase in slit length, roughly between 0.20 and 0.39 s. The reason for lowering the graph at the end was the same as in the previous experiment with different vertical lengths: I thought that the longer the slit, the faster it would fall.

Note. Black graph: Predictive data; Red graph: Real data; Horizontal axis: Prototype slit length; Vertical axis: Flight time.

The interdisciplinary aspects (i.e., engineering and science relevance) in students' descriptive and explanatory use of mathematical and statistical models are shown in Table 3.

Table 3

Interdisciplinary Aspects in Students' Use of Mathematical and Statistical Models

Roles	**Mathematical model**	**#**	**Statistical model**	**#**
Descriptive means	Engineering relevance		N/A	
	Understanding trends (signal) in the data in relation to the prototypes	2		
Exploratory means	Engineering relevance		Engineering relevance	
	Explaining the data trends (signal) using information from the prototypes	20	Explaining reasons for varying flight times (noise) using information from the prototypes	5
	Science relevance		Science relevance	
	Explaining the data trends (signal) using informal scientific knowledge of air resistance	4	Explaining reasons for varying flight times (noise) using informal scientific knowledge of air resistance	1
	Explaining the data trends (signal) based on observations from the experiment	1		

Note. Statements that applied to more than one category were counted.

Twenty-five students used mathematical and/or statistical models to make engineering and/or science-related predictions. Regarding descriptive means, the students used the

prototypical context to read trends (signals) in the graphs. They described the length of the wings and the weight of the prototype by replacing the line graph variable from the flight time with the weight of the prototype. As for the explanatory means, the students used information from prototypes (i.e., volume and weight remained the same, balance of form, and centre of gravity), informal scientific knowledge of air resistance, and observations from the experiment with prototypes to explain the data trends (signals) and reasons for varying flight times (noise). For instance, Hata explained data trends by relating a prototype's weight, shape, and informal scientific knowledge of air resistance to the slope of a graph (Figure 3). In both mathematical and statistical models, the explanatory means were more explicitly related to engineering (particularly prototypes) and science than the descriptive means.

Discussion and Concluding Remarks

This study addressed some aspects of Year 4 students' data-driven predictions in an integrated STEM context, taking the descriptive and explanatory functions of models as perspectives. We discuss two findings regarding this research question.

First, all the participating students consciously or unconsciously used mathematical or statistical models for descriptive and explanatory purposes (Table 2). In line with Oslington et al. (2023), mathematical and statistical models helped students recognise the patterns and structural features of data to support their predictive reasoning. However, more than half of the students tended to use mathematical models only for descriptive and explanatory purposes. This may be because the primary school students involved in this study were mainly exposed to mathematical models rather than statistical models in their everyday mathematics lessons (MEXT, 2018), and they have difficulty understanding complex variations (e.g., Watson et al., 2023). The fixed form of the line graph representation used for predictions might have further triggered the generation of a mathematical model (cf. Oslington et al., 2023; Watson et al., 2023), thereby indicating that more research is necessary to examine students' use of the statistical model in data-driven predictions.

Second, 80% of students connected mathematical or statistical models to other STEM subjects (Table 3). As seen in the case of Hata (Figure 3), mathematical or statistical models, when connected with models in other STEM subjects (i.e., prototypes and scientific models), describe the characteristics of the data and also provide explanatory power for the causal relationships behind the characteristics of the data (i.e., the reasons why flight time varied). Through these explanations with models, students' own scientific hypotheses and interdisciplinary knowledge—linking mathematics and engineering or science—are constructed (Figures 3 and 4), leading to the development of epistemic knowledge (Tytler, 2020). This finding provides evidence of the need for multiple models in demonstrating explanatory power (Ärlebäck & Doerr, 2020) and an example of a reasonable explanation in an integrated STEM context (Baptista et al., 2023).

The findings of this study that data-driven predictions in an integrated STEM context, including IDDM (Kawakami & Saeki, in press), could contribute to the development of STEM content knowledge in justifying predictions as well as their application. On the one hand, data-driven predictions contributed to other STEM subjects by encouraging the generation of students' scientific hypotheses and interdisciplinary knowledge through the process of using models as explanatory tools in prediction. On the other hand, other STEM subjects contributed to data-driven predictions by making sense of the data and models and strengthening the validity of predictions. These findings extend English's (2023) results, which revealed students' application of multidisciplinary knowledge, and advance the role of mathematics/statistics in STEM education as well as in the mathematics curriculum as more than just a service subject (e.g., ACARA, 2022; English, 2016; MEXT, 2018). These findings are based on a case study of the activity of a single prediction in the Seed Dispersal Task. A future step is to analyse in

detail the teacher's role in promoting student predictions as well as the relationship between the development of students' predictions through the task and how they use mathematical and statistical models.

Acknowledgements

This study was partially supported by JSPS KAKENHI Grant Numbers JP21K02513 and JP21K02553. We thank Shohei Chiba for conducting the classroom practice and providing class details. Ethics approval H23–0111 was granted by Utsunomiya University, and principal, teacher, parents, and students in the class gave informed consent.

References

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2022). *The Australian Curriculum (v9.0)*. ACARA. https://v9.australiancurriculum.edu.au/

Aridor, K., Dvir, M., Tsybulsky, D., & Ben-Zvi, D. (2023). Living the DReaM: The interrelations between statistical, scientific and nature of science uncertainty articulations through citizen science. *Instructional Science*, *51*(5), 729–762. https://doi.org/10.1007/s11251-023-09626-8

Ärlebäck, J. B., & Doerr, H. M. (2020). Moving beyond descriptive models: Research issues for design and implementation. *Advances in Research in Mathematics Education*, *17*, 5–20. https://doi.org/gjgsmz

Baptista, M., Jacinto, H., & Martins, I. (2023). What is a good explanation in integrated STEM education?. *ZDM Mathematics Education*, *55*(7), 1255–1268. https://doi.org/10.1007/s11858-023-01517-z

English, L. (2016). Advancing mathematics education research within a STEM environment. In K. Makar, S. Dole, J. Višňovská, M. Goos, A. Bennison, & K. Fry (Eds.), *Research in mathematics education in Australasia 2012–2015* (pp. 353–371). Springer.

English, L. (2023). Multidisciplinary modelling in a sixth-grade tsunami investigation. *International Journal of Science and Mathematics Education*, *21*(Suppl. 1), 41–65. https://doi.org/mbs5

Fitzallen, N., Wright, S., & Watson, J. (2019). Focusing on data: Year 5 students making STEM connections. *Journal of Research in STEM Education*, *5*(1), 1–19. https://doi.org/10.51355/jstem.2019.60

Geiger, V., Gal, I., & Graven, M. (2023). The connections between citizenship education and mathematics education. *ZDM Mathematics Education*, *55*(5), 923–940. https://doi.org/10.1007/s11858-023-01521-3

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.

Innabi, H., Marton, F., & Emanuelsson, J. (2023). Sustainable learning of statistics. In G. F. Burrill, L. de Oliveria Souza, & E. Reston (Eds.), *Research on reasoning with data and statistical thinking: International perspectives* (pp. 279–302). Springer.

Kawakami, T. (2023a). *Research on the learning and teaching of data-driven modelling in school mathematics* [Doctoral dissertation, in Japanese]. Hyogo University of Teacher Education.

Kawakami, T. (2023b). A triplet of data/context, mathematical model, and statistical model: Conceptualising data-driven modelling in school mathematics. In P. Drijvers, C. Csapodi, H. Palmér, K. Gosztonyi, & E. Kónya (Eds.), *Proceedings of the 13th congress of the European Society for Research in Mathematics Education* (pp. 1243–1250). Budapest: Hungary, Alfréd Rényi Institute of Mathematics and ERME.

Kawakami, T., & Saeki, A. (in press). 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 understandings*. Springer.

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.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2018). *Elementary school teaching guide for the Japanese course of study (notified in 2017): Mathematics (Grade 1–6)* [in Japanese]. https://www.mext.go.jp/content/20211102-mxt_kyoiku02-100002607_04.pdf

Oslington, G., Mulligan, J., & Van Bergen, P. (2023). Shifts in students' predictive reasoning from data tables in years 3 and 4. *Mathematics Education Research Journal*. https://doi.org/mb63

Tytler, R. (2020). STEM education for the twenty-first century. In J. Anderson, & Y. Li (Eds.), *Integrated approaches to STEM education* (pp. 21–43). Springer.

Watson, J., Wright, S., Fitzallen, N., & Kelly, B. (2023). Consolidating understanding of variation as part of STEM: Experimenting with plant growth. *Mathematics Education Research Journal*, *35*(4), 961–999. https://doi.org/10.1007/s13394-022-00421-1