PRINCIPLES OF TASK DESIGN TO PROMOTE INFORMAL STATISTICAL INFERENCE IN SECONDARY SCHOOL: COVARIATIONAL REASONING USING MULTIVARIATE DATA

Koki Hosoda
Graduate School of Comprehensive Human Sciences, University of Tsukuba, Japan
hosoda.koki.wy@alumni.tsukuba.ac.jp

In today's data-driven society, statistical inference is required for decision-making and it is essential to develop the relevant skills in school education. This study aimed to develop the task design principles that allow students to experience informal statistical inference by exploring relationships between variables in multivariate data. Therefore, it built on task design principles from a theoretical perspective based on covariational reasoning. Furthermore, a tentative task, "Decathlon Data Problem," was devised based on these principles and clarified guidelines for teachers to promote informal statistical inference through covariational reasoning with multivariate data at the secondary school level.

INTRODUCTION

In statistics education research, much attention has been given to informal statistical inference (ISI) instruction, with an emphasis on understanding the concepts of variability and uncertainty (Makar & Rubin, 2018). More recent studies have focused on ISI with multivariate data. For example, Kazak et al. (2021) analyzed the characteristics of ISI based on students' reports on data modeling. Their study revealed that students used a variety of statistical measures and visualizations to account for variability. It also identified that they did not fully recognize uncertainty when making inferences based on associations between variables.

However, insufficient research has been conducted on task design to help enhance students' understanding of relationships between variables and their ability to consider uncertainty, and to promote ISI. This study, focusing on the process of covariational reasoning, discusses task design principles that promote ISI with multivariate data. Covariational reasoning is essential when assessing the relationships between and changes in bivariate data (e.g., Cobb et al., 2003) and has been crucial to statistical reasoning for exploring multivariate data (Gil & Gibbs, 2017). Therefore, referring to research on covariational reasoning, it would be possible to design an assignment in which students experience an activity that considers uncertainty while exploring the relationships between variables in multivariate data. This study constructed task design principles based on theoretical perspectives put forward in the research community and designed a tentative task for this overall goal. At a later stage, the authors and teachers will jointly re-examine the principles and the task from practical perspectives, enact the task in classrooms, and refine these principles and the task based on the results.

THEORETICAL FRAMEWORK

Informal Statistical Inference (ISI)

As defined by Makar and Rubin (2009), ISI is defined as the process of probabilistic generalization based on available data without the use of formal statistical inference methods, such as hypothesis testing and interval estimation. Their study characterized three necessary elements of ISI. The first one is "generalization beyond the data at hand," which is the process of examining trends and variations in an unknown sample or population from the data available about them and formulating hypotheses. The second element is "the use of data evidence of generalization," which utilizes trends, relationships, and characteristics of context-related data as evidence for inferences. The third element is "the expression of uncertainty using probabilistic (non-deterministic) language" and refers to using the language of uncertainty because inferences from the sample data at hand to the population involve uncertainty. Probabilistic language can suggest that the conclusions obtained are not valid in all cases and need not necessarily be quantified, such as a p-value. Recent studies have indicated the need for ISI studies with correlational and multivariate data. Dierdorp et al. (2011) pointed out that, even though many real-world problems involve relationships between two or more quantitative variables, at the school level little is known about how to infer beyond correlations. Kazak et al. (2021) also indicated that existing studies on ISI tend to examine the process using only data with small samples. This study

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

focuses on developing a teacher's instructional guide to promoting ISI with multivariate data among students to overcome this limitation.

Covariational Reasoning

Covariational reasoning (or reasoning about the association between two variables) identifies or interprets the relationship between two variables and is applied in various fields such as psychology, science, and mathematics (Garfield et al., 2008). Characteristics of learning about covariational reasoning include exploring the variability of individual variables, focusing on the strength and shape of the relationships between variables, and generalizing and explaining those relationships (Cobb et al., 2003). Furthermore, Gil and Gibbs (2017) presented the modeling approach to covariational reasoning in the context of multivariate data analysis. Their study suggests the modeling features for covariational reasoning as follows: the first is the "dimension of the association," which aims to deepen the understanding of multivariate data. Moreover, in "views of the association," the global (aggregate) view and the local view are necessary. Thus, students need to be supported in understanding the individual values of the data (or some of them) and the distribution and structure of the data. Based on these views, relationship modeling is performed using text, visualization, and mathematical representations. Furthermore, "concluding from the association" requires predictions, inferences, and explanations based on findings of prior practices. Here, the students must also focus on distinguishing signals from noise in the data (Engel & Sedlmeier, 2011). The idea of signal and noise is a helpful statistical perspective that can be used to recognize variability.

Principles of Task Design to Promote ISI through Covariational Reasoning with Multivariate Data

This study derived three principles of task design for promoting ISI with multivariate data based on Makar and Rubin's (2009) three elements of ISI and previous studies on covariational reasoning (see Table 1). Task design in this study included the development of the problem and the teacher's instructional guidance related to solving that problem.

Table 1. The principles of task design in this study

Principle	Description of the principle
Principle 1	Posing problems that require of investigating the structure of multivariate data supplied by third-party providers to make sense of open data.
Principle 2	Planning the practice with dynamic statistical technology tools to facilitate inferences from the data based on visualization and assessment of relationships between variables.
Principle 3	Encouraging students to reflect on their own conclusions and reasoning processes to express uncertainty while recognizing the signal and noise in the data.

The first principle refers to the problem presented to students. Based on a review of previous studies, Gil and Gibbs (2017) considered the modeling of covariational reasoning with open data and multivariate data to be important in the statistical education curriculum. Their study also explained that covariational rationale is necessary to explore and identify the hidden structure of the data, which helps to deepen their understanding. Therefore, statistical educators and teachers need to pose problems for students to explore the structure of multivariate data supplied by third-party providers.

The second principle concerns the practice of visualizing and evaluating the relationships between variables. To promote covariational reasoning, students should be relieved of the complex calculation of the data and allowed to extend the modeling from the data (Gil & Gibbs, 2017). Specifically, TinkerPlots software (Konold & Miller, 2011) and Excel can be used to calculate correlation coefficients or explore cross and sliced scatterplots (Cobb et al., 2003; Konold, 2002). Instruction that relates group comparisons of univariate distributions with cross and sliced scatterplots to covariational reasoning supports students' understanding of the correlations. Therefore, teachers need to plan instruction with dynamic statistical technology tools to facilitate inferences from the data based on visualization and assessment of relationships between variables.

The third principle refers to the teacher's guidance on reflecting upon conclusions and reasoning processes and considering uncertainty. Because it is assumed that students are usually unfamiliar with distinguishing signals from noise in data and drawing sound conclusions, active intervention by the teacher and additional problem positioning may be necessary. For example, this could include describing the confidence and uncertainty level for another or future set of data in the same context. Further, in the context of the problem, the role of variables that are not present in the data should be considered. Therefore, teachers need to encourage students to recognize uncertainty by reflecting on their conclusions and reasoning processes.

DEVELOPING THE TASK: DECATHLON DATA PROBLEM

The task developed in this study is an updated version of Hosoda's (in press) task based on task design principles (see Figure 1). This task uses data from the decathlon of the 2021 Tokyo Olympics dataset (n = 23, variables for the 10 event performances and the overall score)—data that can be downloaded from the International Association of Athletics Federations (https://worldathletics.org/).

> The "King of Athletes" is a decathlon in athletics, wherein the participants' ranking is determined by the sum of their scores in ten different events. Here, you are working as a member of the team's training department, analyzing what events are important for a player to be successful in the international decathlon competition.
> Task 1. To understand what events are important. What are the events that seemed to influence the overall score? Identify these events based on the data from the 2021 Tokyo Olympics. Further, explain the reasons for your conclusion.
> Task 2. Consider what you need to reach better conclusions in Task 1.

Figure 1. The task of the "Decathlon Data Problem"

For the relationship between the performance in each event and the overall score in Task 1, correlations for the 110-meter hurdles (rounded off to two decimal places, -0.83), 100-meter run (-0.66), and long jump (0.65) have the highest values. Thus, one example solution is to identify the event based on the correlation coefficient. Another solution could discuss the events based on the center of the distribution, using a cross option or a sliced scatterplot, as displayed in Figure 2. These two ideas are helpful as a stage before the normal instruction of scatterplots because they are based on the analysis of the mean and median with the univariate distribution. In Task 2, the goal is to critically consider the conclusions and reasoning process in Task 1. For example, students should highlight that the conclusions of the analysis are limited to the data from the Tokyo Olympics, thus to a specific competition; therefore, it is difficult to draw conclusions about data from other competitions in a deterministic way. This second task also facilitates the understanding of uncertainty through the consideration of variables that are not present in the data used (e.g., climatic conditions).

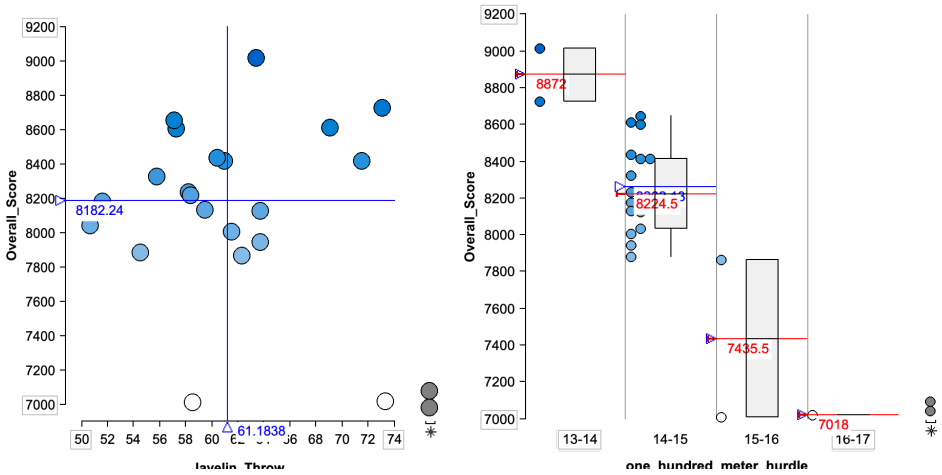

Figure 2. Scatter plots of the cross option (*left*) and slice option (*right*)

DISCUSSION AND CONCLUSION

The following is a relationship among the three principles for task design and the designed task. The relationship among the three principles for the task design and the designed task can be presented as follows: the first principle posed problems with investigating the structure of multivariate data supplied by third-party providers to make sense of open data. In Figure 1, the context was designed with the aim of allowing students to explore the relationship between the variables of performance in each event and the overall score based on the open Tokyo Olympics data. The second principle was to plan instruction using dynamic statistical technology tools to facilitate inferences from the data available based on visualization and assessment of relationships between variables. As shown in Figure 1, the practice was planned to enable students to explore and make inferences using the analysis methods based on correlation coefficients and univariate distributions of the relationships between the variables of each event performance and the overall score. The third principle was to encourage students to reflect on their own conclusions and reasoning processes to consider uncertainty while recognizing the signal and noise in the data. Accordingly, it is suggested that teachers encourage students to reflect on previous practices and understand uncertainty in view of the sample variability and the context related to the phenomenon in the problem. In this study, the relationship between ISI and covariational reasoning was reconsidered, clarifying theoretical aspects of teaching ISI, including correlations. The limitation of this study is that it is still only a theoretical consideration by a single community of researchers. In the future, a team consisting of the authors and faculty members will conduct practical consideration and elaborate further on the task design principles and the task.

ACKNOWLEDGMENT

This study was supported by JSPS KAKENHI Grant Number JP20J10388.

REFERENCES

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and instruction*, *21*(1), 1–78. https://doi.org/10.1207/S1532690XCI2101_1

Dierdorp, A., Bakker, A., Eijkelhof, H., & Maanen, J. V. (2011). Authentic practices as contexts for learning to draw inferences beyond correlated data. *Mathematical Thinking and Learning, 13*(1–2), 132–151. https://doi.org/10.1080/10986065.2011.538294

Engel, J., & Sedlmeier, P. (2011). Correlation and regression in the training of teachers. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-Challenges for teaching and teacher education. A joint ICMI/IASE study: The 18th ICMI Study* (pp. 247–258). Springer https://doi.org/10.1007/978-94-007-1131-0_25

Garfield, J., Ben-Zvi, D., Chance, B., Medina, E., Roseth, C., & Zieffler, A. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer. https://doi.org/10.1007/978-1-4020-8383-9_14

Gil, E., & Gibbs, A. L. (2017). Promoting modeling and covariational reasoning among secondary school students in the context of big data. *Statistics Education Research Journal, 16*(2), 163–190. https://doi.org/10.52041/serj.v16i2.189

Hosoda, K. (In press). Task design to foster informal statistical inference in upper secondary school: Focusing on correlation with multivariate data. *Proceedings of the 12th Congress of the European Society for Research in Mathematics Education (CERME12)*.

Kazak, S., Fujita, T., & Turmo, M. P. (2021). Students' informal statistical inferences through data modeling with a large multivariate dataset. *Mathematical Thinking and Learning*. Advance online publication. https://doi.org/10.1080/10986065.2021.1922857

Konold, C. (2002). Alternatives to scatterplots. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics*, *Cape Town, South Africa*. International Statistical Institute. https://www.stat.auckland.ac.nz/~iase/publications/1/7f5_kono.pdf

Konold, C., & Miller, C. (2011). *TinkerPlots* (Version 2.3) [Computer software]. Key Curriculum Press.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82–105. https://doi.org/10.52041/serj.v8i1.457

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: Cham. https://doi.org/10.1007/978-3-319-66195-7_8