The Mathematical Education. Vol. 62, No. 2, 289-302
https://doi.org/10.7468/mathedu.2023.62.2.289

Print ISSN 1225-1380
Online ISSN 2287-9633

Research Article

Guidelines for big data projects in artificial intelligence mathematics education[1]

Lee, Junghwa[1] · Han, Chaereen[2] · Lim, Woong[3]*

[1]Graduate student, Yonsei University

[2]Adjunct professor, Yonsei University

[3]Professor, Yonsei University

*Corresponding Author: Woong Lim (woonglim@yonsei.ac.kr)

ABSTRACT

In today's digital information society, student knowledge and skills to analyze big data and make informed decisions have become an important goal of school mathematics. Integrating big data statistical projects with digital technologies in high school <Artificial Intelligence> mathematics courses has the potential to provide students with a learning experience of high impact that can develop these essential skills. This paper proposes a set of guidelines for designing effective big data statistical project-based tasks and evaluates the tasks in the artificial intelligence mathematics textbook against these criteria. The proposed guidelines recommend that projects should: (1) align knowledge and skills with the national school mathematics curriculum; (2) use preprocessed massive datasets; (3) employ data scientists' problem-solving methods; (4) encourage decision-making; (5) leverage technological tools; and (6) promote collaborative learning. The findings indicate that few textbooks fully align with these guidelines, with most failing to incorporate elements corresponding to Guideline 2 in their project tasks. In addition, most tasks in the textbooks overlook or omit data preprocessing, either by using smaller datasets or by using big data without any form of preprocessing. This can potentially result in misconceptions among students regarding the nature of big data. Furthermore, this paper discusses the relevant mathematical knowledge and skills necessary for artificial intelligence, as well as the potential benefits and pedagogical considerations associated with integrating technology into big data tasks. This research sheds light on teaching mathematical concepts with machine learning algorithms and the effective use of technology tools in big data education.

Key words: artificial intelligence, AI mathematics textbook, big data tasks, statistical literacy, use of technology

인공지능 수학 교육을 위한 빅데이터 프로젝트 과제 가이드라인

이정화[1] · 한채린[2] · 임웅[3]*

[1]연세대학교 대학원생 · [2]연세대학교 겸임교수 · [3]연세대학교 교수

*교신저자: 임웅 (woonglim@yonsei.ac.kr)

초록

지식정보사회의 비약적인 발전에 힘입어 빅데이터를 분석하여 가치있는 결과물을 도출하고 유용한 정보를 추출하는 역량이 학교 수학의 주요 목표 중 하나로 급부상하고 있다. 고등학교 수학 진로 선택 과목 중 하나인 <인공지능 수학>은 디지털 기술을 활용한 통계 프로젝트를 통해 빅데이터에 기반한 새로운 통계 교육의 기회를 제공할 수 있다. 이 연구에서는 효과적인 빅데이터 통계 프로젝트 기반 과제를 설계하기 위한 일련의 가이드라인을 제안하고, 이 기준에 따라 5종의 인공지능 수학 교과서에 실린 최적화 단원 과제들을 평가하였다. 인공지능 수학 교과에서 빅데이터 통계 프로젝트 과제를 설계 시

[1] The data in this study were initially compiled and analyzed as part of the first author's thesis work.

Received April 18, 2023 **Revised** May 04, 2023 **Accepted** May 10, 2023
2000 Mathematics Subject Classification : 97U20

고려하도록 도출된 가이드라인은 다음과 같다: (1) 지식과 기술을 국가 학교 수학 교육과정에 맞추고, (2) 전처리된 대규모 데이터 세트를 사용하며, (3) 데이터 과학자의 문제 해결 방법을 사용하고, (4) 의사 결정을 장려하며, (5) 공학도구를 활용하고, (6) 협업 학습을 촉진한다. 분석 결과에 따르면 가이드라인에 완전히 부합하는 과제는 드물었고, 특히 대부분의 교과서에서 가이드라인 2에 해당하는 요소를 프로젝트 과제에서 통합하지 못하고 있는 것으로 나타났다. 또한 소규모 데이터 세트나 빅데이터를 전처리 없이 직접 사용하는 경우가 많아 학생들의 빅데이터의 개념에 대한 오해를 불러일으킬 것이 우려된다. 본 연구에서는 결과를 토대로 인공지능에 필요한 관련 수학 지식과 기술을 밝히고, 이것이 빅데이터 과제에 통합될 때 얻을 수 있는 잠재적 이점과 교육적 고려사항에 대해 논의하였다. 이 연구는 수학적 개념과 머신러닝 알고리즘과의 연계 및 빅데이터를 사용하는 통계 교육에서의 효과적인 공학적 도구 사용에 대한 통찰을 제공하고자 하였다.

주요어: 인공지능, 인공지능 수학 교과서, 빅 데이터 과제, 통계적 소양, 기술 활용

Introduction

In the digital information society, big data has emerged as a new paradigm of making decisions serving as a catalyst for uncovering insights and trends that would have otherwise remained hidden within the vast quantities of information (Lee & Kim, 2015). In response, statistical literacy or the ability to analyze data, derive valuable outputs, and make decisions has become an important goal of school mathematics since it equips students with the necessary skills for data collection, interpretation, reasoning, and decision-making (Ministry of Education, 2015). With the rapid emergence of big data and artificial intelligence, however, educational and curricular contexts have become increasingly complex, necessitating a reform in traditional statistical literacy to effectively address the challenges they present (Lee et al., 2021).

The shift to big data is not just a change in data types; there has also been a significant change in the nature of data -- big data is dynamic and its value can change over time as new analysis techniques are used (Bargagliotti et al., 2020). The change in the nature of data has led to changes in analysis methods, and the use of artificial intelligence and machine learning models helps analyze large amounts of data (LeCun et al., 2015). Mathematics education researchers argue that with the advent of machine learning and artificial intelligence, changes in the type and nature of data, as well as analysis methods, should serve as the foundation for data-based statistics education (Lee et al., 2021).

As our future becomes increasingly data-driven, it is important to define and develop a new form of statistical literacy that incorporates big data thinking and artificial intelligence while making related mathematical concepts more accessible to students (Heo, 2020). While traditional statistics courses have primarily focused on imparting content knowledge, we understand the importance of implementing a new pedagogical approach that emphasizes data literacy with artificial intelligence in a broader sense, and entails the use of digital technologies for engaging students in meaningful learning opportunities (see Ben-Zvi, 2000) through big data statistical projects within an "Artificial Intelligence Mathematics" (AI Mathematics) course.

In light of the growing importance of teaching artificial intelligence, mathematics, and data literacy together, current research trends in AI Mathematics and related digital technologies in mathematics education can be classified into two distinct categories. The first category focuses on developing effective teaching methods that support student learning in using mathematical concepts foundational for writing AI algorithms. For example, Park and Kim (2022) designed and implemented a teaching program on simple linear regression analysis and classification to promote understanding of the basic principles of mathematics in AI in a high school setting. They found that students' recognition of the role and value of mathematics in AI had increased. Ko (2020) underscored the importance of developing big data thinking in students and enabling them to recognize the value and usefulness of statistics and mathematics, as well as gain a basic understanding of artificial intelligence, such as machine learning models. The second category involves analyzing curricular material in AI Mathematics, examining the content, learning elements, and differences in computational thinking processes (Han, 2022; Kim et al., 2021; Kwon et al., 2021). These studies suggest that incorporating AI into mathematics education has the potential to improve students' understanding of mathematics and artificial intelligence. However, more research is needed to design and evaluate curricula that effectively integrate big data and AI, fostering statistical literacy and big data thinking in students. Specifically, research on curriculum and pedagogy required to develop statistical literacy for big data and AI is required. This involves creating innovative curricular materials for teaching high school students how to analyze big data, understand the principles of AI, and engage in statistical tasks within the context of artificial intelligence.

This paper aims to (1) propose guidelines for designing big data statistical tasks for high school students and (2) assess the current AI Mathematics textbooks against these criteria. In doing so, this paper seeks to identify gaps and areas that require improvement in the current curricular material.

Theoretical backgrounds

Statistical literacy and big data thinking

Gal (2002) provided a definition of statistical literacy as the ability to interpret statistical data, engage in critical thinking, and make statistical inferences. Statistical thinking can be divided into knowledge elements and disposition elements. Knowledge elements encompass basic skills such as comprehension, effective communication, and the performance of simple computations. Statistical knowledge includes concepts such as mean, median, mode, variance, correlation, standard deviation, sample mean, and population mean. Mathematical knowledge is supplementary to statistical knowledge and consists of concepts like calculus, equations, and inequalities. Contextual knowledge pertains to the application of statistical knowledge to real-life situations. Disposition elements refer to critical attitudes and beliefs towards statistical content and topics, including a critical stance that scrutinizes statistical data, analysis, and interpretations to ensure accuracy and prevent biases. Critical thinking or questioning in this context involves reflecting on the proper understanding of statistical information and assessing the reliability of statistical data.

Garfield and Ben-Zvi (2004, 2009) provided more nuanced definitions for statistical thinking and statistical literacy. Statistical thinking concerns the rationale and methods used to explore statistical data, including the recognition of bias in variability, the selection of appropriate visualization and summary techniques, and choosing the appropriate timing and methods of data analysis. In contrast, statistical literacy refers to the basic skills needed to understand statistical information, such as statistical concepts, terms, and symbols.

Critical questioning in statistical literacy includes evaluating the reliability of the data collection process, determining the rationality of the analysis method, and using appropriate statistical knowledge for data analysis (Gal, 2004; Gould, 2017). In big data thinking, critical questioning extends beyond statistical thinking to include the design or selection of suitable algorithms developed with artificial intelligence. Big data thinking uses statistical thinking to analyze vast datasets and construct predictive models using artificial intelligence, such as machine learning (Secchi, 2018). Furthermore, it involves critical thinking that acknowledges the limitations of big data for prediction.

Drawing from the nuanced juxtaposition between statistical and big data thinking paradigms (Garfield & Ben-Zvi, 2009; Secchi, 2018), this study operationalizes the concept of big data thinking as the process of analyzing big data, which distinctly deviates from traditional statistical data, and building predictive models through AI models such as machine learning.

Statistical data vs. big data

The differences between statistical data in textbooks and big data can be summarized in four main points. First, the size of the data sets differs. Textbook statistical data typically consists of 30 to 50 pieces of numerical data about real-life situations, while big data is characterized by a massive amount of data exceeding the volume of conventional databases (Bühlmann & van de Geer, 2018).

Second, the datasets in textbook tasks and big data tasks have different relationships (Secchi, 2018). Textbook statistical tasks involve a sample, which is a subset of the population, and primarily focus on estimating the population mean using the sample. Big data tasks, on the other hand, analyze training data to build a predictive model, often evaluating the model's accuracy using testing data. Consequently, the four datasets (sample vs. population and training vs. testing) present distinct relationships for students to understand.

Third, big data is observational statistical data collected without a specific design purpose (Kim & Cho, 2013). Textbook statistical projects involve setting variables expected to have a correlation and collecting data related to those variables. In contrast, big data is collected without a specific purpose. To address problems presented in big data, statistical analysis is required to directly identify the inherent meaning of the data (Galeano & Peña, 2019). A hypothesis is then established based on this meaning, and the data are analyzed to verify the hypothesis.

Fourth, the purpose of data interpretation differs between statistical tasks and big data. The revised 2015 education curriculum's achievement standards for probability and statistics state that students should estimate the population mean and interpret the results with a small sample (Ministry of Education, 2015). Therefore, textbook statistical tasks involve estimating the population mean using a small sample through statistical analysis. In contrast, big data analyzes the entire population and uses variables with significant correlations to find the meaning and value of the data (Galeano & Peña, 2019; Kim & Cho, 2013). The goal is to build a prediction model based on the results of the big data analysis, using a separate set of testing data from the training data to measure the accuracy of the prediction algorithm and construct the optimal prediction model.

In sum, the literature indicates the need for guidance on understanding the preprocessing steps associated with big data and distinguishing it from conventional statistical data and emerging pedagogical considerations regarding the loose (versus rigorous) definition and uses of big data within the school context.

Statistical problem solving and big data statistical project tasks

The description of statistical problem solving varies in the literature, but it generally involves setting the problem, planning data collection, collecting data, analysis, and solving the problem. The PPDAC model (Mackay & Oldford, 2000) is a classic model for statistical problem solving and consists of five stages: Problem, Plan, Data, Analysis, and Conclusion. Built from the PPDAC model, Wild and Pfannkuch (1999) summarized the problem-solving process of statisticians and students. However, this model could be challenging to apply to school-level statistics education (Kim & Jeon, 2021). The American Statistical Association proposed an alternative model for statistical problem solving in the GAISE report (Franklin et al., 2007). Franklin et al. (2007) classified the statistical problem-solving process into four stages suitable for school-level statistics education. In the problem-setting stage, the problem is defined, and questions regarding data are formed. In the data collection stage, a plan is made to collect appropriate data to solve the problem, and the plan is executed to collect the necessary data. In the data analysis stage, appropriate visual and numerical methods are selected to analyze the data. Finally, the data is interpreted based on the results of the data analysis, and the problem is solved.

Considering the aforementioned stages and the four data science life cycle models in Lee and Han (2020, see Table 1 for a summary), the big data statistical project task in this paper focuses on the data processing and predictive modeling stages, similar to that of a data scientist conducting a project for problem-solving.

Table 1. Data science lifecycle model from Lee & Han (2020)

Stage	Description
Problem identification	Understand the problem accurately.
Data collection	Collect data that can help solve the problem and check if additional data is needed.
Data preprocessing	Examine data through statistical analysis and preprocess for machine learning models.
Predictive model building and execution	Build and execute the predictive model for problem-solving using machine learning.
Communication	Share the predictive model built with machine learning through data visualization and improve with feedback.

More specifically, we propose the type of task to find significant parts of big data through statistical analysis, build a predictive model through artificial intelligence to predict the output value corresponding to the input value, and measure the accuracy of the model in the following manner (see Table 2): First, it is essential to have an accurate understanding of the problem. Second, it involves collecting data that helps solve the problem and assessing whether more data is needed. Third, statistical thinking is used to analyze the significant characteristics of the data and to process the data for machine learning. Further, hypotheses are formed using the significant characteristics of the data. Fourth, a predictive model is made through machine learning and implemented to solve the problem. Finally, the results of the predictive model are presented through data visualization to receive feedback and improve the predictive model.

Table 2. Comparing statistical problem solving process versus data science lifecycle models

Statistical problem solving process	Data science lifecycle model	Descriptions
Problem setting	Problem identification	Setting the problem to be solved
Data collection	Data collection	Checking the data collection plan and execution
Data analysis	Data preprocessing	Checking the size of the collected data
	Predictive modeling	Understanding the relationship between data used in the problem-solving process
Interpreting results	Communication and maintenance	Evaluating the problem-solving of the task

Methods

Proposing guidelines. This study uses the document analysis method to establish guidelines for designing big data statistical tasks and assess the current AI Mathematics textbooks based on those guidelines. <AI Mathematics> is an elective course in the 2015 revised curriculum. Currently, there are five AI Mathematics textbooks in Korea, and the object of analysis of this study is the project tasks of the optimization chapter in the five textbooks. The primary objective of this study is not so much to select or score textbooks as to assess alignment and analyze tasks; therefore, the names of the publishers for the textbooks examined are not disclosed.

The AI Mathematics subject of the 2015 revised curriculum consists of several content areas, including AI and Big Data, Text Data Processing, Image Data Processing, Prediction and Optimization, and Exploring AI and Mathematics. The Prediction and Optimization section, which is the focus of text analysis, includes the following curriculum standards:

[12인수04-01] Analyze data to find the probability of an event occurring and use it to make predictions.
[12인수04-02] Analyze data to determine the probability of an event occurring and use it for decision-making.
[12인수04-03] Understand loss functions and find optimized trend lines.
[12인수04-04] Understand the gradient descent method and explain how artificial intelligence learns for optimized predictions.

The focus of the analysis is on whether and to what extent the project tasks align with the big data statistical project guidelines proposed in this study. Our document analysis process involves *finding*, *selecting*, *evaluating*, and *synthesizing* data from documents (Bowen, 2009). In the *finding* stage, we searched for the kind of disciplinary practice related to statistical problem-solving (Franklin et al., 2007) and big data statistical project work (Lee and Han, 2020). We then *selected* principles for carrying out statistical project tasks with big data based on these findings. Next, we *checked* the principles of designing statistical project tasks against the principles of good mathematics task design (Lee et al, 2017; see Table 3). The good mathematics tasks framework has the following principles: First, good mathematics tasks should be low-threshold, high-ceiling tasks, thus engaging students at various levels, reflective of the multifaceted nature of mathematics. Second, the task should help form mathematical connections, allowing students to recognize the relationships between mathematical concepts and see them as a connected whole. Third, the task should be in the form of an open task that allows for a variety of representations and conjectures, enabling students to discover mathematical relationships while solving the task and presenting them as tentative propositions and hypotheses. Fourth, the task should provide students with the right to make decisions during the solving process, motivating them to actively participate in the task. Fifth, the task should provide opportunities for exploration, allowing students to make conjectures among mathematical objects and justify or criticize them while solving the task. Sixth, the task should promote collaboration, enabling students to reflect on his or her peers' thinking and facilitating a diverse inquiry. Finally, we derived the guidelines for designing statistical project tasks in the final *synthesis* stage by examining whether the guidelines adhere to the good task framework as well as the professional practice of data scientists' problem-solving (see Table 3).

Table 3. Six principles of good mathematical task design (Lee et al., 2017) and big data statistical project task designing guidelines

Principles of good math task design	Big data statistical project-based task development guidelines
Provide low floor high ceiling tasks	Align with the 2015 revised curriculum
	Provide preprocessed big data
Experience mathematical connections	Use the problem-solving practice endorsed by data scientists and data science educators
Use various expressions, conjectures, and solutions	Perform future predictions
Encourage students' decision-making and self-directedness.	
Provide opportunities for exploration.	Use technology tools
Promote peer interaction and collaboration.	Promote collaborative learning.

Evaluating textbook tasks. Using these guidelines, we reviewed the project tasks in the optimization section of the five AI Mathematics textbooks in Korea. Since the proposed guidelines are appropriate for project tasks, we focused our evaluation on project tasks in the optimization section of five AI Mathematics textbooks. The focus of the analysis was on whether and to what extent the project tasks adhere to the big data statistical project guidelines developed in this study.

To assess whether Guideline 1 was met, we examined if the tasks included questions that directly addressed mathematical ideas emerging during the investigation process. As these questions pertain to mathematical content required within the curriculum, we referenced the 2015 revised curriculum for our evaluation. For Guideline 2, we assessed the tasks using two criteria: (1) the presence or absence of a process for collecting massive data, and (2) the incorporation of data preprocessing to enhance prediction accuracy for the collected big data. For Guideline 3, we compared the questions presented in project tasks with the components of the data science lifecycle model. The task sequence in most AI Mathematics textbooks typically consists of: data collection for problem-solving, organization and interpretation of collected data, problem-solving using AI, and organization of research findings. This task sequence closely aligns with the data science lifecycle model. For Guideline 4, we considered the guideline met if the tasks encompassed problem-solving using AI or future prediction utilizing an AI algorithm based on training data. For Guideline 5, we determined whether tasks encouraged the application of technology tools for data analysis or the execution of prediction algorithms through technology tools. Finally, for Guideline 6, we checked if tasks were presented as group activities with shared roles and responsibilities.

Results

Presented here is a comprehensive set of guidelines for designing big data statistical project tasks. These are intended to enhance students' mathematical learning experiences and contribute to their understanding of statistical analysis within the context of big data:

· Guideline 1: The project should be in line with the national curriculum and pedagogical guidelines, ensuring it is appropriate for high school students with various mathematical proficiencies. If necessary, the entry threshold for big data statistical projects may be adjusted. When presenting mathematical concepts used in big data and machine learning, the project should stay within the curriculum's scope. The focus should be on building future prediction models using various machine learning algorithms. Introduce mathematical concepts beyond the curriculum in such a way that students can intuitively use technology tools or other methods.

· Guideline 2: The project should include preprocessing of large data sets, acknowledging that statistics are key in understanding uncertainty within big data and predicting future trends. Using statistical tasks with limited sample data, as often seen in conventional statistics textbooks, might not effectively address uncertainties common in today's information societies. It's important to introduce students to large, properly preprocessed data, if needed, to minimize confusion for those unfamiliar with handling big data. However, the importance of allowing students to experience data preprocessing should not be overlooked in favor of convenience when designing big data statistical tasks.

· Guideline 3: The project should adhere to the actual problem-solving process endorsed by data scientists and data science educators, incorporating mathematical concepts related to each stage of the data science lifecycle model. This approach fosters a understanding of connected mathematics and allows students to experience professional mathematics and data science practices, developing authentic problem-solving skills.

· Guideline 4: The project should emphasize future prediction elements through artificial intelligence, such as machine learning models. Building predictive models using various machine learning algorithms in big data tasks promotes decision-making and self-directedness (i.e., Kang et al., 2021). Promoting statistical investigations and encouraging the use of multiple problem-solving strategies can provide students with a deeper understanding of statistics as data science.

· Guideline 5: The project should take advantage of technology tools for problem-solving in big data tasks, enabling students to explore the statistical domain, establish connections between various representational systems of big data, and concentrate on essential statistical exploration. Technology tools also facilitate handling real data and visual recognition of variability, connecting statistics to real-life situations.

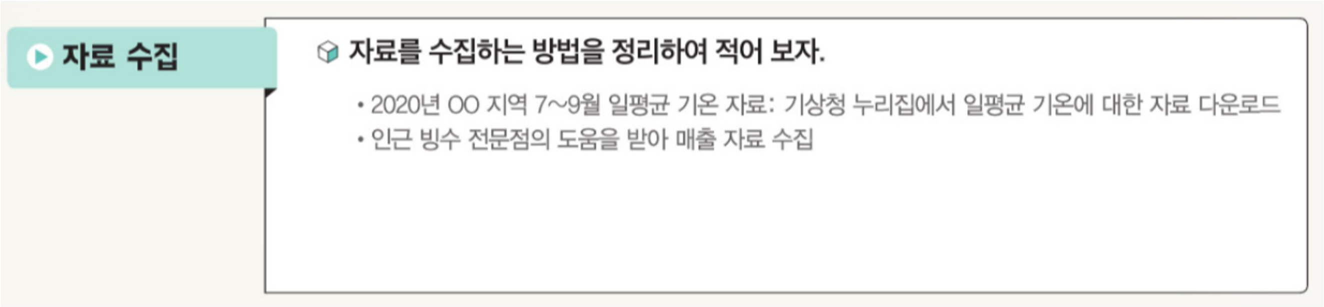
· Guideline 6: The project should foster collaborative learning, providing students with opportunities to share their mathematical ideas and expand their understanding by engaging with others' perspectives. Integrating collaboration into big data statistical projects and promoting effective communication during problem-solving can enhance students' statistical thinking. In the era of artificial intelligence, the importance of communication and collaboration skills is increasing, highlighting the need for group projects.

Based on the analysis of tasks in the five AI Mathematics textbooks (see Table 4) regarding their alignment with the guidelines, we report that the majority of project tasks did not include the use of (or the step to create) preprocessed big data (Guideline 2), technology tools (Guideline 5), or collaborative learning (Guideline 6) (Lee et al., 2017).

Table 4. Analysis of AI Mathematics textbook project tasks against the guidelines

Textbook	Guideline 1	Guideline 2	Guideline 3	Guideline 4	Guideline 5	Guideline 6
A	○	╳	○	○	╳	○
B	○	╳	○	○	╳	○
C	○	▲	○	○	○	╳
D	-	-	-	-	-	-
E	○	▲	○	╳	╳	╳

Concerning Guideline 2, the project tasks in AI textbooks were found to have students to collect big data directly during the problem-solving process, rather than providing (or asking to create) preprocessed data. Tasks were designed to produce solutions without big data collection or to directly apply AI for problem-solving without preprocessing big data. For instance, one task by Textbook C (see Figure 1) encourages big data collection during the data collection phase but does not necessarily include a data preprocessing step.

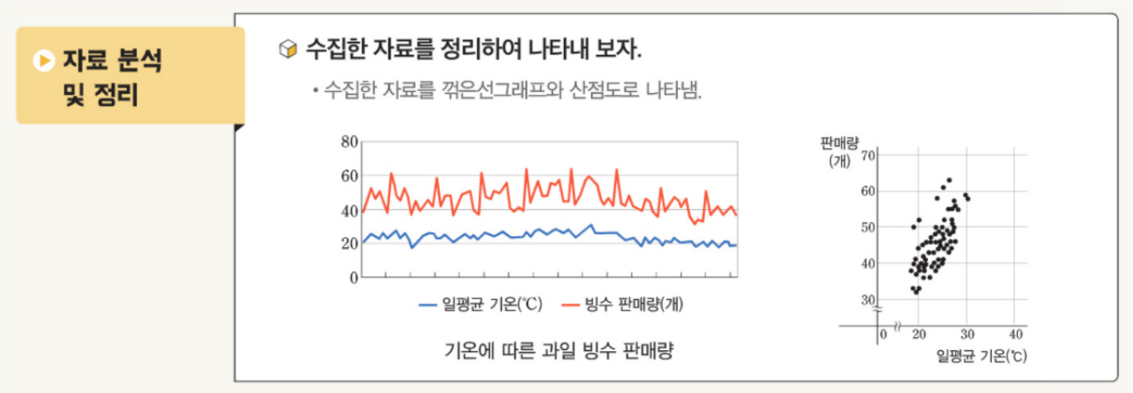

자료 수집

자료를 수집하는 방법을 정리하여 적어 보자.

- 2020년 OO 지역 7~9월 일평균 기온 자료: 기상청 누리집에서 일평균 기온에 대한 자료 다운로드
- 인근 빙수 전문점의 도움을 받아 매출 자료 수집

Figure 1. Data collection stage of AI Textbook C (p.138)

We also note that the dataset under consideration might not meet the typical size criteria often attributed to big data. Nevertheless, this dataset, comprising at least 180 data points, possesses a relative scale substantial enough to be classified as big data. Viewed from the standpoint of 'relationships' within statistical data, this dataset is more of big data than others because the dataset is used to analyze the correlation between two variables in the entire dataset in marked contrast to the conventional approach of employing a sample to estimate the population mean (see Figure 2). Regarding the purpose of data collection, the dataset may have not been generated to discover the shaved ice sales according to the temperature with a specfic design purpose. Taken in concert, we concluded this dataset indeed exemplified the core attributes of big data.



자료 분석 및 정리

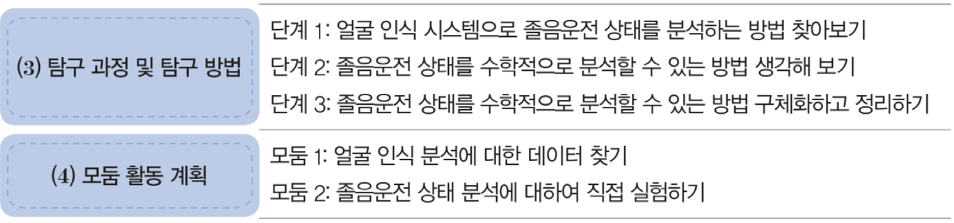
수집한 자료를 정리하여 나타내 보자.

• 수집한 자료를 꺾은선그래프와 산점도로 나타냄.

Figure 2. Data analysis stage of AI Textbook C (p. 138)

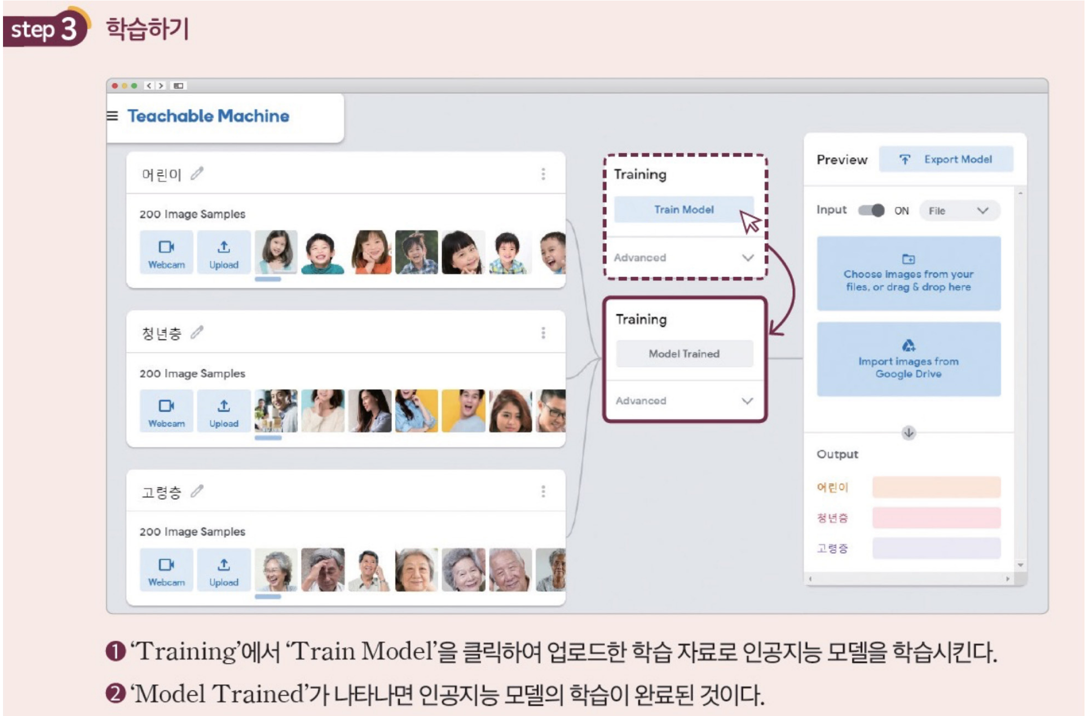
However, such tasks may lead to misconception about big data among students. First, if data preprocessing is not performed to preserve data characteristics, extreme data points may lead to inaccurate predictions. Second, students may misunderstand the characteristics of big (or raw, unprocessed) data during the data collection process for problem-solving -- they may perceive little difference between big data and “clean datasets” in conventional statistics textbooks. This finding confirms Lee and Rim (2021) and Lee et al. (2021), highlighting that big data and statistics education, in its present form, provides limited opportunities for students to investigate or interact with data as an important component of problem-solving. Instead, the textbooks seem to use data mostly as a contextual medium for elucidating complex concepts in statistics. Research (i.e., Lee et al., 2021) indicates teachers should emphasize that big data, as purposeless observational statistics, requires statistical analysis to discover its inherent patterns and meanings. Choi (2017) adds that teachers should support student for preprocessing the big data and instruct students to recognize the differences between clean data and big data.

Concerning Guideline 5, only one textbook presented a task where technology tools were used for data analysis and future prediction. Most textbooks focused on analyzing mathematical concepts necessary for designing algorithms to solve problems rather than predicting the future using technology. Some textbooks (see Figure 3a/3b) suggested the use of technology tools for data analysis but actually omitted the data analysis procedure in which students need technology. This finding indicates a notable absence of pedagogy necessitating students to execute the designed prediction algorithm using technology, despite the plethora of research supporting the integration of technology in AI Mathematics (i.e., Kim & Jeon, 2021) as well as within reformed statistics education (Lee & Rim, 2021; Lee et al., 2021).



(3) 탐구 과정 및 탐구 방법	단계 1: 얼굴 인식 시스템으로 졸음운전 상태를 분석하는 방법 찾아보기 단계 2: 졸음운전 상태를 수학적으로 분석할 수 있는 방법 생각해 보기 단계 3: 졸음운전 상태를 수학적으로 분석할 수 있는 방법 구체화하고 정리하기
(4) 모둠 활동 계획	모둠 1: 얼굴 인식 분석에 대한 데이터 찾기 모둠 2: 졸음운전 상태 분석에 대하여 직접 실험하기

Figure 3a. Example of project-based tasks with investigation and group activity from Textbook A (p. 125)


step 3 학습하기

❶ 'Training'에서 'Train Model'을 클릭하여 업로드한 학습 자료로 인공지능 모델을 학습시킨다.

❷ 'Model Trained'가 나타나면 인공지능 모델의 학습이 완료된 것이다.

Figure 3b. Example of project-based tasks from Textbook B (p. 125)

Lastly, the textbooks did not address collaborative learning corresponding to Guideline 6. Collaborative learning fosters self-directed task-solving among students and improves the quality of social relationships as a result of positive interdependence (Johnson & Johnson, 2008). In the context of big data analysis, collaboration is essential for several reasons. First, big data analysis often involves complex and multi-faceted problems, requiring a diverse range of skills and knowledge. Collaborative learning enables students to leverage their individual strengths, creating a synergy that allows for more effective problem-solving (see a project-based curriculum in Davidson et al., 2019 for example). Second, the interdisciplinary nature of big data analysis necessitates collaboration among students with diverse backgrounds. Collaborative learning prepares students for real-world scenarios where they will need to communicate effectively and work together with professionals from various fields to tackle challenges associated with using big data. In addition, collaboration promotes critical thinking and creativity by exposing students to different perspectives and ideas. This can lead to creative and diverse solutions and in-depth analysis, as students learn to challenge assumptions and consider alternative approaches to problem-solving (Gal, 2002; Kim, 2019).

Discussion

In this study, we identified two primary concerns regarding the use of AI-based mathematics textbooks in secondary education: the alignment of appropriate mathematical knowledge (Guideline 1) and the use of technology tools (Guideline 5) within the tasks. We argue that issues surrounding Guidelines 2, 3, and 4 are relatively uncomplicated, and as the field gains more foundation in ontologies and the theoretical framework of teaching and learning, the curricular challenges may be resolved. The issue related to Guideline 6 extends beyond mathematics or big data education, making it more relevant to broader national reform agendas.

Further, we posit that Guidelines 1 and 5 are closely related, as the use of technology has the potential to increase students' engagement with advanced mathematical concepts in AI Mathematics, which typically fall outside the purview of the national mathematics curriculum. By examining key high school mathematics concepts in several machine learning algorithms, we seek to demonstrate that the majority of concepts covered by these algorithms align with the scope of the 2015 revised mathematics curriculum. However, we also find that there are some mathematics concepts, such as those related to the decision tree algorithm or cost function (i.e., support vector machine), that are currently not listed in the national school mathematics curriculum but could be considered appropriate for the secondary classroom. In order to address this concern, we argue leveraging technology tools can support student learning in big data by facilitating intuitive understanding of complex concepts and correcting student misconceptions if any.

Machine learning algorithms and secondary mathematics content

In AI-based mathematics education, the discussion regarding the extent to which the concepts of traditional mathematics curricula are covered is increasingly important, yet it has not been widely addressed as a significant issue in the literature. Here, we aim to explore key high school mathematics concepts within several machine learning algorithms, as part of the conversation about the scope of mathematical concepts in AI-based secondary mathematics education.

To predict the future through machine learning algorithms, it is necessary to design a prediction model using machine learning algorithms. The machine learning algorithms considered here is within the scope of the 2015 revised mathematics curriculum. Table 5 shows the secondary mathematics content for the machine learning algorithms.

Table 5. Secondary mathematics content for machine learning algorithm

Learning machine algorithms			Secondary mathematics contents
Supervised learning	Regression	- Linear regression	Vectors, correlation, linear regression (least squares method)
	Classification	- Decision trees	-
		- Support vector machines (in 2D space)	Vectors (dot product, projection), distance between point and line
		- K-nearest neighbors	Euclidean distance
		- Logistic regression	Graph shape of logistic function (extreme values, maxima, minima)
Unsupervised learning	Clustering	- K-means	Euclidean distance, 3D coordinate system, centroids

First, linear regression analysis uses the concept of correlation, which is related to the weights of each variable, and uses mean squared error and the least squares method to express the value of the dependent variable for the independent variable as a single linear equation. In simple linear regression analysis, there is only one independent variable and one dependent variable, so understanding the principle of mean squared error and least squares method through graphs may not exceed secondary mathematics content. For the task of using linear regression analysis for prediction, the task therefore should present with simple linear regression analysis without using multivariable concepts. However, the task should use technology tools for the actual calculation of the least squares method because the partial derivative concept is required in the process of estimating $y = \beta_0 + \beta_1 x$ using the least squares method in simple linear regression analysis. Additionally, the task should engage students in studying the relationship between the principle of the least squares method and the statistical variance.

Secondly, decision tree algorithms use inequalities to make binary decisions. However, many components of the algorithm, such as nodes, branches, depth, entropy, and the Gini index, are not secondary mathematics content in the national curriculum (Ko, 2020). Nevertheless, these concepts and related formulas are regular (and appropriate for a secondary classroom) curricular topics in discrete mathematics. This highlights a scenario where AI Mathematics has some concepts that may be relatively straightforward to comprehend, but are not currently included in the national school mathematics curriculum. Similarly, support vector machine models use advanced mathematics, including the Lagrange multiplier method. However, in a 2D space, these models can determine the optimal decision boundary, represented by the equation $ax + by = c$, using vector operations such as dot products and vector projections. Using technology, students can calculate the maximum margin as a function of two variables, typically expressed as $f(a, b)$. Regarding the concept of kernel functions in support vector machines, related mathematics does go beyond what's typically covered in school-level mathematics, students can still gain an understanding of kernels as a mapping, illustrated through transformations using quadratic and exponential functions.

Third, the K-nearest neighbors (KNN) algorithm classifies new data by measuring its proximity to already classified data points. This depends on the distance between points in geometric coordinate space, or other advanced mathematical distance metrics. Therefore, the task should propose writing the classification algorithm using the Euclidean distance between data points, instead of using other distance measures. Similarly, the K-means algorithm, also relying on the principle of Euclidean distance, creates clusters by assigning data points to the nearest centroid. This process, which involves calculating the mean of data points within a cluster, uses mathematical concepts taught in secondary mathematics.

Fourth, logistic regression analysis defines the logistic function that best classifies data by determining the constants of the logistic function. Thus, understanding the shape of the logistic regression function (i.e., sigmoid function) is important to understand the principle of logistic regression analysis. The understanding of the sigmoid function requires caculus, but the task can use a step function instead. As for sensitivity and specificity to evaluate the accuracy of the constructed algorithm, the understanding of Bayes' theorem is necessary, which is beyond the scope of the school mathematics curriculum. Alternatively, sensitivity and specificity can be taught as the ratio of the number of actual data points with a value of 1 among the predicted data points with a value of 1, as a concept of algorithm accuracy measurement. In big data thinking, the more information available, the more rational the decision-making process and the higher the prediction accuracy of the model, which is related to abductive reasoning.

Using technology tools in big data tasks

Expected benefits. Leveraging technology tools can enhance the learning process in big data by facilitating intuitive understanding (see Yeo, 2021) of complex concepts. For instance, introducing matrices through technology tools enables students to grasp matrix representation without formally learning the topic. This can be achieved by using spreadsheets to represent the weighted sum of training data after teaching perceptron algorithm principles, allowing students to intuitively comprehend matrices and matrix multiplication. Moreover, technology tools can address common student misconceptions in the current curriculum. High school students may struggle with judging correlation strength, relying on a visual density or a regression line slope rather than the regression line's proximity. In big data analysis and regression algorithm design, correlation plays a key role in determining the weights of variables on outcomes. Technological tools can help rectify misconceptions by visualizing big data, calculating actual correlations between variables and outcomes, and facilitating their comparison. For instance, spreadsheet functions are capable of calculating Pearson's correlation coefficient and the slopes of linear regression lines, as well as visualizing data through scatter plots effectively.

Pedagogical considerations. To optimize the use of technology tools in teaching big data and machine learning, teachers should focus on conveying difficult mathematical concepts intuitively and providing students with opportunities to explore mathematical thinking and reasoning. The primary aim of big data tasks is to help students grasp machine learning procedures and principles on a macroscopic level by solving problems and understanding the underlying statistical and mathematical concepts. As machine learning principles involve repeated calculations, technology tools prove invaluable in managing these tasks. However, these tasks do not necessitate teaching specific programming languages like Python or R, as the main objective is to foster big data thinking rather than focusing on coding skills. Out of five AI Mathematics textbooks related to statistical tasks in a study (Kwon et al., 2021), three tasks do not implement technology tools. However, teachers should aim for a balanced approach, integrating technology tools and traditional teaching methods to effectively promote big data thinking and dispel potential statistical misconceptions -- the emphasis needs to be on improving students' understanding of AI processes and the mathematics behind the algorithms with scaffolding such as pre-written code.

Implications

This study contributes to the ongoing conversation about the integration of AI and big data into secondary mathematics education. The proposed guidelines for designing effective big data statistical project-based tasks and evaluating AI Mathematics textbook tasks against these criteria serve as a starting point for improving AI Mathematics education. The results of our analysis indicate that most textbooks do not meet

Guideline 2 in their project tasks, which suggests the need for further research on creating specific methods for creating big data statistical project tasks that engage students in a big data preprocessing. Moreover, teaching mathematical concepts within the secondary mathematics curriculum can sometimes lead to oversimplified or contextually unrelated tasks. Therefore, it's important to conduct research on pedagogical approaches for teaching mathematical concepts that fall outside the curriculum's scope, particularly in the context of artificial intelligence mathematics. In that sense, we need more research to explore instructional strategies that use technology tools to support such pedagogy. While technology tools have the potential to represent mathematical concepts intuitively, their current use in AI Mathematics textbooks is primarily limited to problem definition, computable abstraction, and result interpretation. Teachers should harness these tools to facilitate a deeper understanding of the mathematical concepts and principles inherent in AI, rather than employing them superficially.

The appropriateness of the proposed guidelines for assessing the AI Mathematics course, especially the optimization unit, remains open to discussion. The course's stated objective is to understand how mathematics can be applied to resolve a variety of real-life problems through artificial intelligence, acknowledge the significance of mathematics, and cultivate the skills necessary for future society (Ministry of Education, 2020). However, our guidelines focus on problem-solving using artificial intelligence algorithms, akin to the role of an AI scientist. Furthermore, the study's textbook task analysis was limited to *whether* the textbooks adhere to the guidelines. Additionally, the present study has developed and evaluated guidelines that, in essence, function as curricular standards. The term 'standards', which could alternatively be described as criteria, principles, policies, or best practices, may not be easily written and scrutinized by a select group of researchers. Undoubtedly, a thorough theoretical underpinning is a prerequisite, as often expected in educational projects of national scale.

In light of these limitations, it is incumbent upon future research to revise these guidelines to better align with the national vision for the curriculum, with input from key stakeholders in mathematics education, including leading experts, scholars, and policymakers. With that in mind, the findings of this study should be viewed as insights provided by individual researchers, contributing to the collective effort of informing the national committee. It is through this collaborative endeavor that we can facilitate the development of an effective mathematics curriculum designed to meet the ever-evolving needs of our students for the future.

References

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. A. (2020). *Pre-K-12 guidelines for assessment and instruction in statistics education Ⅱ: A framework for statistics and data science education.* American Statistical Association.

Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical Thinking and Learning, 2*(1&2), 127–155. https://doi.org/10.1207/S15327833MTL0202_6

Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27-40. https://doi.org/10.3316/QRJ0902027

Bühlmann, P., & van de Geer, S. (2018). Statistics for big data: A perspective. *Statistics & Probability Letters, 136*, 37-41. https://doi.org/10.1016/j.spl.2018.02.016

Choi, D. (2017). Problems of big data analysis education and their solutions. *Journal of the Korea Convergence Society, 8*(12), 265-274. https://doi.org/10.15207/JKCS.2017.8.12.265

Davidson, M. A., Dewey, C. M., & Fleming, A. E. (2019). Teaching communication in a statistical collaboration course: A feasible, project-based, multimodal curriculum. *The American Statistician, 73*, 61–69. https://doi.org/10.1080/00031305.2018.1448890

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-k-12 curriculum framework*. American Statistical Association.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70*(1), 1-51. https://doi.org/10.1111/j.1751-5823.2002.tb00336.x

Gal, I. (2004). Statistical literacy: Meanings, components, responsibilities. In J. B. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 47–78). Kluwer Academic Publishers.

Galeano, P., & Peña, D. (2019). Data science, big data and statistics. *TEST, 28*, 289–329. https://doi.org/10.1007/s11749-019-00651-9

Garfield, J., & Ben-Zvi, D. (2004). Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 397-409). Kluwer Academic Publishers.

Garfield, J., & Ben-Zvi, D. (2009). Helping students develop statistical reasoning: Implementing a statistical reasoning learning environment. *Teaching Statistics, 31*(30), 72–77. https://doi.org/10.1111/j.1467-9639.2009.00363.x

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal, 16*(1), 22-25. https://doi.org/10.52041/serj.v16i1.209

Han, S. H. (2022). New mathematical competence for <Artificial Intelligence Mathematics>: A focus on digital competence. *The Journal of Educational Research in Mathematics, 32*(1), 1-22. https://doi.org/10.29275/jerm.2022.32.1.1

Heo, N. G. (2020). Analysis of secondary mathematics knowledge for AI learning through the AI related R&E Program. *The Journal of Learner-Centered Curriculum and Instruction, 20*(16), 673-689. https://doi.org/10.22251/jlcci.2020.20.16.673

Johnson, D. W., & Johnson, R. T. (2008). Social interdependence theory and cooperative learning: The teacher's role. In R. Gillies, A. Ashman, & J. Terwel (Eds.), *The teacher's role in implementing cooperative learning in the classroom. Computer-supported collaborative learning.* Springer. https://doi.org/10.1007/978-0-387-70892-8_1

Kang, H. R., Lim, C. L., & Cho, H. H. (2021). A study on coding mathematics curriculum and teaching methods that converges school mathematics and school informatics. *The Mathematical Education, 60*(4), 467-491. https://doi.org/10.7468/mathedu.2021.60.4.4671

Kim, C., & Jeon, Y. J. (2021). The core concepts of mathematics for AI and an analysis of mathematical contents in the <AI Mathematics> textbook. *Journal of the Korean School Mathematics, 24*(4), 391-405. https://doi.org/10.30807/ksms.2021.24.4.004

Kim, H. (2019). A study on the direction of future education in the AI era. *The Journal of Future Education, 9*(4), 1-15. https://doi.org/10.26734/JFE.2019.09.04.01

Kim, Y., & Cho K. (2013). Big data and statistics. *Journal of the Korean Data and Information Science Society, 24*(5), 959-974. https://doi.org/10.7465/jkdi.2013.24.5.959

Ko, H. K. (2020). A study on development of school mathematics contents for artificial intelligence (AI) capability. *Journal of the Korean School Mathematics Society, 23*(2), 223-237. https://doi.org/10.30807/ksms.2020.23.2.003

Kwon, O. N., Lee, K., Oh, S. J., & Park, J. S. (2021). An analysis of 'related learning elements' reflected in <Artificial Intelligence Mathematics> textbooks. *Communications of Mathematical Education, 35*(4), 445-473. https://doi.org/10.7468/jksmee.2021.35.4.445

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lee, D. H., Go, E. S., Kwon, S. I., Kim, D. W., Kim, Y., Park, J. H., Gu, N., & Lee, K. H. (2017). *Designing and implementing tasks for inquiry in school mathematics* (Korea Foundation for the Advancement of Science & Creativity BD 18020001). Retrieved from https://askmath.kofac.re.kr/board.do?menuPos=15&menuPos=15&act=detail&idx=12271&searchValue1=title&skinSearchValue2=&searchKeyword=%EC%A2%8B%EC%9D%80&pageIndex=2

Lee, E. H., & Kim, W. K. (2015) A comparative analysis on research trends of statistics education between Korea and overseas. *The Mathematical Education, 54*(3), 241-259. https://doi.org/10.7468/mathedu.2015.54.3.241

Lee, H. W., & Han, S. H. (2020). An analysis of data science curriculum in Korea. *Journal of the Korean Society for Library and Information Science, 54*(1), 365-385. https://doi.org/10.4275/KSLIS.2020.54.1.365

Lee, J., & Rim, H. (2021). Analysis of <Probability and Statistics> textbooks on statistical problem-solving process and statistical literacy. *The Korean School Mathematics Society, 24*(2), 191-216. https://doi.org/10.30807/ksms.2021.24.2.002

Lee, K. H., Yoo, Y., & Tak, B. (2021). Towards data-driven statistics education: An exploration of restructuring the mathematics curriculum. *The Korea Society of Educational Studies in Mathematics, 23*(3), 361-386. https://doi.org/10.29275/sm.2021.09.23.3.361

MacKay, R. J., & Oldford, R. W. (2000). Scientific method, statistical method and the speed of light. *Statistical Science*, *15*(3), 254-278. https://doi.org/10.1214/ss/1009212817

Ministry of Education (2015). *Mathematics curriculum* (#2015-74 supplement 8). Ministry of Education.

Ministry of Education (2020). *Mathematics curriculum* (# 2020-236 supplement 8). Ministry of Education.

Park, J. I., & Kim, S. B. (2022). The development and effect analysis of customized artificial intelligence and mathematics convergence program for vocational high schools. *The Journal of Korean Association of Computer Education*, 25(3), 39-47. http://doi.org/10.32431/kace.2022.25.3.004.

Secchi, P. (2018). On the role of statistics in the era of big data: A call for a debate. *Statistics & Probability Letters, 136*, 10-14. https://doi.org/10.1016/j.spl.2018.02.041

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

Yeo, S. (2021). Semiotic mediation through technology: The case of fraction reasoning. *The Mathematical Education, 60*(1), 1-19. https://doi.org/10.7468/mathedu.2021.60.1.1

Authors Information

Junghwa Lee, Yonsei University, Graduate student, 1st Author.
ORCID: https://orcid.org/0009-0009-8186-7127

Chaereen Han, Yonsei University, Adjunct professor, Co-author.
ORCID: https://orcid.org/0000-0001-7956-3049

Woong Lim, Yonsei University, Professor, Corresponding Author.
ORCID: https://orcid.org/0000-0002-4329-952X