INVESTIGATIVE QUESTIONS WITH SECONDARY DATA: CHARACTERIZING HIGH SCHOOL STUDENTS' QUESTIONS AND THE ROLE OF DATA VISUALIZATION IN REFINEMENT

HYEJIN JUN
Seoul National University
mathedu86@snu.ac.kr

KYEONG-HWA LEE
Seoul National University
khmath@snu.ac.kr

ABSTRACT

We investigated how high school students formulate and refine investigative questions when conducting a statistical investigation with secondary data. The data consisted of students' written activity reports and email-based, post-interview responses collected after a seven-session instructional sequence in which CODAP served as the primary tool for multivariate data analysis. We distinguished initiating investigative questions (IIQs) from analysis-phase investigative questions (AIQs) implied in students' analysis plans and characterized both sets across five analytical components: variables, clarity of population, intent, feasibility of drawing conclusions from the data, and global view of data. We then used thematic analysis to examine how data visualization appeared to be involved at points where IIQ-to-AIQ refinement was evident. Compared with IIQs, AIQs more often included a greater number of clearer variables and took forms that were more feasible for drawing conclusions from the given dataset. Across the three episodes analyzed, the representational and exploratory functions of data visualizations appeared to support refinement by helping students operationalize everyday terms, narrow populations, anticipate interpretable relationships, and set aside uninformative variables. This study offers classroom-based empirical insights into investigative questions in the context of secondary data and into the potential role of data visualization in their refinement.

Keywords: *Statistics education research; Statistical investigation; Investigative questions; Secondary data; Data visualization; High school students*

1. INTRODUCTION

Recent school curricula show a growing emphasis on statistical investigations, an emphasis that has expanded to include key practices from the field of data science (Fielding et al., 2025; Lee et al., 2022). With the rise of data science and the widespread availability of public data, secondary data (i.e., data collected by others) have likely become the most common type of data students encounter outside the classroom (Bargagliotti & Gould, 2022). For students living in a data-driven era, the ability to use such data to pose questions proactively and to examine and analyze data critically is becoming increasingly important (Bargagliotti et al., 2020; Engel, 2017). Secondary data can provide rich opportunities to investigate relationships among variables, variability, and uncertainty (Lee & Wilkerson, 2018), while also offering contexts that many students find interesting and that can provoke meaningful investigative questions (Bargagliotti & Gould, 2022). Despite these advantages, working with secondary data can be challenging. Such data often take the form of large multivariate datasets for which the definitions and measurement of variables, as well as the purposes and methods of data collection, are predetermined. Recent studies have begun to explore the use of public and multivariate datasets in diverse contexts (e.g., Kazak et al., 2023; Zapata-Cardona, 2025).

Statistical investigation is a nonlinear and dynamic process in which investigators repeatedly revisit phases; it is often initiated by the posing of an investigative question (Lee et al., 2022). Investigative questions are central because they initiate the investigation and guide the overall process, thereby serving as a key indicator of the quality of inquiry (Frischemeier & Biehler, 2018). Such questions are often initially posed in vague and general forms and may be refined over the course of an investigation

Statistics Education Research Journal, 25(1). https://doi.org/10.52041/serj.820

(Arnold & Franklin, 2021). Prior studies have explored instructional supports that help students to refine investigative questions, including teacher feedback, peer negotiation, and instructional designs involving tasks and prompts (e.g., Allmond & Makar, 2010; Leavy & Frischemeier, 2022). Leavy and Frischemeier (2022) argued that the data themselves should guide the formulation and refinement of investigative questions. In particular, questions should fit the available dataset in terms of its size and structure, whether it contains adequate variability, and whether it invites a global view of the data. Yet, in the context of working with large, multivariate, secondary datasets, relatively little is known about the extent to which students' consideration of the data themselves—and of the requirements and constraints they encounter when analyzing those data—is reflected in the formulation and refinement of their investigative questions.

Digital tools are integral to students' engagement with data in statistical investigations and are essential for working with large, multivariate datasets, enabling students to make sense of the data, generate visualizations, and explore information relevant to their intentions and questions (Biehler et al., 2024). Interactive visualization tools such as the Common Online Data Analysis Platform (CODAP; The Concord Consortium, 2014) can support students in drawing evidence-based conclusions through active data exploration (Sutherland & Ridgway, 2017). Data scientists also use data visualization as an essential tool in their work (Bolch & Crippen, 2022). They use data visualization for various purposes, one of which is to gain insights from data while keeping their research questions in mind. Drawing on data scientists' practices, Lee et al. (2022) proposed a data investigation framework that positions data exploration and visualization as a distinct phase of investigation and conceptually suggests that investigative questions may be refined during this phase. However, this possibility has not yet been examined empirically in classroom settings.

To address the research gaps identified above, we focused on how the requirements and constraints encountered when planning data analysis are reflected in high school students' investigative questions when they engage in statistical investigations with secondary data, as well as on the potential of data visualizations in this context. The purpose of this study was to characterize the investigative questions formulated in this setting and to explore how data visualizations were involved in investigative question refinement. To this end, we conceptualized two types of investigative questions: (1) initiating investigative questions (IIQs) that frame the overall investigation and (2) analysis-phase investigative questions (AIQs) that emerge during analysis planning (see Section 3.4). We analyzed the characteristics of IIQs and AIQs using five analytical components associated with investigative questions (e.g., Arnold, 2013): variables, clarity of population, intent, feasibility of drawing conclusions from the data, and global view of data. We then examined the role of data visualizations using CODAP at points where qualitative improvements were evident in comparisons between IIQs and AIQs. The research questions were as follows.

(1) In a statistical investigation using secondary data, how are students' IIQs and their associated AIQs characterized across five analytical components?
(2) Within the same activity, what role does data visualization appear to play in students' refinement of IIQs into AIQs?

2. THEORETICAL BACKGROUND

2.1. INVESTIGATIVE QUESTIONS IN STATISTICAL INVESTIGATION

An effective approach to fostering meaningful statistical learning in school mathematics is to engage students informally in statistical investigations that mirror the practices of statisticians or data scientists (Leavy & Frischemeier, 2022; Watson & English, 2017). Prior studies (e.g., Lee et al., 2022; Wild & Pfannkuch, 1999) have proposed related stage-based frameworks for statistical investigation, which commonly portray the process as iterative and dynamic rather than strictly linear. Among these, Lee et al. (2022) extended a data investigation framework by analyzing the practices of data scientists. The framework consists of six stages: framing the problem, considering and gathering data, processing data, exploring and visualizing data, considering models, and communicating and proposing actions. In particular, it separates what is often treated as a single data analysis phase into two stages: data processing and the exploration and visualization of data. This distinction reflects the important roles of data processing and visualization when problems are addressed with large and complex datasets in data science. As Cobb and Moore (1997) noted, "data are not just numbers, they are numbers with a context" (p. 801), suggesting that statistical investigations can be used to address questions grounded in meaningful contexts. Accordingly, attention to the quality of students' investigative questions is warranted.

Drawing on prior work, we outline five components of investigative question quality that attend to question alignment with the constraints and possibilities of a given dataset. First, clarity about the variables of interest—what they are and how they are defined and measured—plays an important role in formulating investigative questions. When working with a given dataset, however, students sometimes overlook variables that are explicitly available in the data and instead frame questions around vaguely specified or even absent variables (e.g., Arnold, 2013; Frischemeier & Biehler, 2018). For example, students might ask "Are boys healthier than girls?" when the dataset only includes specific measures such as body mass index or weekly exercise time.

Second, clarity of the population to which the question refers is a key component of formulating investigative questions. Students may treat a sample as the population (e.g., "these students") or invoke overly broad groups (e.g., "people," "boys") that extend beyond what the dataset can warrant (Arnold, 2013). Arnold (2013) noted that even small wording shifts can change the referent from an individual case to the sample, or from the sample to a broader population. Providing guidance on how the data were collected and whom they represent (e.g., sampling context, inclusion criteria, key descriptors) can help students articulate a warranted target population and align the question with the scope of the dataset (Bargagliotti et al., 2020).

Third, the intent of the investigative question is an important consideration because it directly relates to the selection of variables. According to Graham (2006), the intent of an investigative question can be categorized into three types: summary, comparison, and relationship. Summary questions aim to describe or summarize key features of a single variable—for example, "What is the typical amount of time elementary students spend on homework each day?" Comparison questions aim to identify differences in an attribute across groups (e.g., comparing homework time by gender), and thus typically involve one categorical variable and one numerical variable. Relationship questions are bivariate questions about paired data that examine how two variables—categorical, discrete, or continuous—measured on the same observational units are associated (Arnold, 2013). Recent guidelines also extend the scope of question intent to include prediction. The GAISE II report (Bargagliotti et al., 2020), for example, recommended that students at Level C—roughly corresponding to high school—be encouraged to formulate questions with predictive elements, reflecting the growing significance of predictive modeling in statistics and data science (e.g., linear regression and random forests; Biehler et al., 2024; James et al., 2013).

Fourth, an investigative question should be answerable using the given data. In statistical investigations, questions that aim to draw inferences about a population—and the conclusions drawn from the investigation—should be grounded in data (Makar & Rubin, 2009). Meeting this fourth component can be particularly difficult when a question involves variables that are not included in the data, or when the available variables are difficult to interpret or to operationalize. In addition, if the available sample is too small, it may not be possible to draw statistically meaningful conclusions (Leavy & Frischemeier, 2022). Questions that are overly broad or vague can also pose challenges. Nevertheless, ambiguity need not be viewed solely as a deficit; it can sometimes serve as a resource for inquiry (Allmond & Makar, 2010). However, ambiguity becomes productive only when students can anticipate how the question can be addressed with the available data and how resulting claims can be warranted as conclusions. Otherwise, conclusions drawn from a statistical investigation may become disconnected from the investigative question (Allmond & Makar, 2010).

Fifth, investigative questions typically focus on the group as a whole, rather than on individual observations. This distinction relates to how data are viewed: a local view emphasizes individual data points, whereas a global view attends to patterns such as trends or measures of central tendency (Ben-Zvi & Arcavi, 2001). Such a shift from a local to a global view is consistent with Bakker and Gravemeijer's (2004) description of students' development from attending to individual observations to conceiving a data distribution as an entity with global characteristics such as center, spread, and shape. The intent of the question is also relevant. For instance, the question "What is the relationship between height and weight?" reflects a global perspective aimed at identifying an overall trend in the data (Arnold, 2013).

Previous research suggests that the quality of investigative questions can be considered across these five analytical components: variables, clarity of population, intent, feasibility of drawing conclusions from the data, and global view of data. These components underpin the analytic framework used in this study (see Table 1 in Section 3.5), which provides a basis for characterizing students' IIQs and their associated AIQs in statistical investigations using secondary data.

2.2. SECONDARY DATA AND INVESTIGATIVE QUESTION FORMULATION

Secondary data refers to data that have been collected by others, which is typically the type of data with which school students work (Arnold, 2013). Most of the data that students encounter in their everyday lives can be classified as secondary data, and it is important that students have opportunities to engage with such data in statistics education (Bargagliotti & Gould, 2022). In recent years, open data platforms such as Kaggle (https://www.kaggle.com/) have made it possible for anyone to access and analyze publicly available datasets for various purposes. Many public datasets are rich and multivariate, comprising numerous variables of different types that can motivate creative investigations and support students in generating meaningful insights (Podworny et al., 2022). Incorporating secondary data into statistics learning enables students to undertake sophisticated statistical investigations using datasets that would be difficult to collect on their own due to their scale or complexity (Arnold, 2013; Hall, 2011). Advances in technology further support students in exploring large, multivariate datasets, helping them to visualize, transform, and interrogate data as they formulate and investigate statistical questions (Biehler et al., 2024).

Although incorporating secondary data into statistics learning offers several advantages, it also presents challenges. One such challenge involves the use of metadata. As data about the data itself, metadata is a defining feature of working with secondary data. Awareness of metadata is closely tied to evaluating the extent to which conclusions drawn from data are valid and generalizable. Therefore, before formulating investigative questions, students are expected to critically examine the metadata—for example, who collected the data, for what purpose, and how—and, for survey data, to consider the original survey question(s), drawing on this information when deciding what investigative question(s) to pose (Bargagliotti & Gould, 2022; Buehring & Grando, 2023). In this sense, survey questions describe how a dataset was generated, whereas students' investigative questions guide how they analyze that dataset (Arnold & Franklin, 2021), somewhat akin to the kind of student engagement with data generation described by Cobb and McClain (2004). Even if students do not collect the data themselves, engaging in discussions about the metadata can help them consider the original purpose for which the data were collected and formulate investigative questions that align with their own interests and goals (Cobb & McClain, 2004).

A second challenge concerns the complexity of secondary data. Such datasets typically include a large number of variables, which may vary in type and, in some cases, prompt students to explore their meaning (James et al., 2013). Moreover, secondary data often involve many observations, within which missing values or anomalies may be found. Examining the dataset in tabular form and engaging in discussions about the number and types of variables, their meanings, the number of observations, and the presence of missing or extreme values can help students develop a deeper understanding of the data (Tukey, 1977). The inherent complexity of secondary data can be mitigated to some extent by removing variables with excessive missing values or those that are difficult for students to understand, and by preprocessing the data into a more tame or manageable form (Kim et al., 2018). It is important for teachers to recognize that students may feel overwhelmed by the complexity of such data to plan instruction accordingly (Bargagliotti & Gould, 2022).

Statisticians and data scientists often use data visualization when working with data, including large multivariate datasets (Bolch & Crippen, 2022; Franke et al., 2016). With advances in visualization tools, students can also explore such datasets through tool-generated representations (Biehler et al., 2024). As Card et al. (1999) noted, data visualization is not merely a picture; it can offer insights for discovery, decision making, and explanation. Visualizations can represent patterns, variability, and relationships in data, and the processes of generating, comparing, and adjusting multiple displays may also help reveal new findings or exceptions (Biehler et al., 2024; Cairo, 2012). Because posing an investigative question aims to extend the base of contextual knowledge through data (Wild & Pfannkuch, 1999), the representational and exploratory functions of visualization may align with students' formulation of investigative questions, in part by making secondary data more accessible.

The forms of visualization that become relevant can vary depending on which variables are included in an investigative question. Scatterplots, trend lines, and least-squares regression lines are typically used to represent relationships between numerical variables, whereas boxplots may be used to summarize distributions of numerical variables or to examine relationships between a numerical and a categorical variable (Wilke, 2019). In exploring students' approaches to relationships among categorical variables, Higgins et al. (2023) reported that students may also generate insights by creating a contingency graph, a variant of the contingency table. Scatterplot matrices and bubble charts are not typically taught as visualization forms in formal school curricula, yet they can be viewed as extensions of scatterplots and are often used when attempting to visualize three or more variables simultaneously (Wilke, 2019). Thus, the type and number of variables included in an investigative question are related to the forms of visualization that are feasible, which may lead students to refine their questions in different ways.

Secondary data may prompt students to attend to multiple variables and explore patterns, differences, and relationships among them. However, the inherent complexity of secondary data can make it difficult for students to formulate and refine investigative questions. The following section reviews prior research on instructional supports for improving and developing students' investigative questions and examines studies that discuss the potential role of data visualization in question refinement.

2.3. DATA VISUALIZATION AND INVESTIGATIVE QUESTION REFINEMENT

Students with limited experience in statistical investigations and a developing understanding of statistical concepts often formulate investigative questions that are vague and general (Arnold, 2013). Formulating and refining such questions is demanding not only because it involves coordinating multiple features of a question, but also because it requires judging what the available data can support. These demands are especially true when working with data collected by others, where issues such as scope and variability constrain possible analyses (Leavy & Frischemeier, 2022). Prior studies have reported that these challenges can limit progress in subsequent stages of data analysis and interpretation and have identified several types of investigative questions that require refinement (e.g., Arnold, 2013; Leavy & Frischemeier, 2022). These include questions with unclear variables, questions involving only a single variable, questions lacking clarity about the population, questions that cannot be answered with data, questions focused on individual observations, and overly closed-ended, yes-or-no questions.

Prior research has examined how students refine the investigative questions they initially pose during statistical investigations into forms that can be answered with data. These studies have designed and implemented instructional supports for refinement, including task sequences, engaging datasets or stimuli that build interest in the data context (e.g., watching a video), opportunities for peer feedback through small-group discussion, and structured tools such as checklists, and have documented how such supports function within the refinement process (Watson & English, 2017; Allmond & Makar, 2010). In prospective teacher contexts, researchers have reported that feedback and structured activities can support improvements in the clarity of variables and populations and in the development of comparison- and relationship-oriented questions (Leavy & Frischemeier, 2022; Frischemeier & Leavy, 2020). At the same time, even in instruction that emphasizes multivariable questions, studies suggest that refinement patterns can vary, including a tendency to remain focused on single-variable questions (Frischemeier & Biehler, 2018).

Data visualization has often been treated primarily as part of the data analysis phase, but recent studies have reported that data scientists—who frequently work with large multivariate datasets—use visualization for purposes broader than communicating results (e.g., Bolch & Crippen, 2022; Lee et al., 2022). In Lee et al.'s (2022) framework for data inquiry, inquiry begins with broader issues related to real-world problems. After these issues are understood and the variability inherent in the context is considered, one or more investigative questions that can be addressed through a statistical approach are posed. Lee and colleagues emphasized that problem reframing occurs throughout the entire inquiry cycle and suggested that visualization and analytic techniques that support insight into the data may be connected to the refinement of investigative questions. Bolch and Crippen (2022) reported that data scientists interpret visualizations through strategies such as reading axes, attending to the meanings of variables, and making comparisons; they further noted that data scientists design visualizations with the research question in mind by clarifying their purpose and making relationships and patterns more salient.

Although studies examining the relationship between students' investigative questions and data visualization in school mathematics contexts remain limited, findings from Laina and Wilkerson (2016) offer implications for how students might revise or refine investigative questions. They found that middle school students experienced tension among competing interpretations for complex narrative visualizations and responded by adjusting the structure of the visualization or revising their interpretations. These findings suggest that visualization may surface tensions between alternative representations—or different readings of the same display—and that supporting students in negotiating such tensions may contribute to the development of more specific and analytically tractable investigative questions.

This study builds on prior perspectives that conceptualize the refinement of investigative questions as a process of qualitative improvement across the statistical investigation cycle (e.g., Leavy & Frischemeier, 2022; Lee et al., 2022). Grounded in this prior view, refinement in the present study is defined as a qualitative improvement in which students' investigative questions that initiate a statistical investigation are reshaped into analytically tractable forms in response to the requirements and constraints of the data analysis phase. In addition, we attend to the affordances provided by the distinctive features and functions of data visualization tools (Watson & Fitzallen, 2015) that may support investigative question refinement. Here, refinement is not limited to improving sentence structure; rather, it involves making questions clearer and more analyzable so that they can guide data analysis toward meaningful conclusions. Such refinement may involve qualitative changes in one or more respects, and these changes may be partial or holistic. For example, refinement may be reflected in an expansion in the number and types of variables included in a question, such as questions that coordinate numerical and categorical variables to broaden the scope of exploration (Frischemeier & Biehler, 2018). These developments may enable students to explore data from multiple perspectives and support their engagement with the statistical investigation process.

3. METHODS

This study adopted a qualitative case study approach (Merriam & Tisdell, 2016), in which a case was conceptualized as a bounded system situated in a real-life context. In this study, the case was bounded by a seven-session sequence of learning activities implemented with nine grade 11 students (16–17 years old) in one high school. The analytic focus was the investigative questions students posed during the sequence. The research process is summarized in Figure 1.

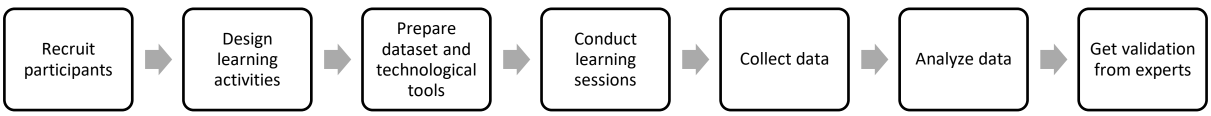


Figure 1. Research process overview

3.1. PARTICIPANTS

The participants in this study were nine high school students (all male), who voluntarily took part in the statistical investigation and data analysis activities. The dataset provided to the students was large and included a wide range of variables. Because working with such a multivariate dataset was likely to be unfamiliar to students and require sustained engagement, we invited students who had completed a Grade 11 Probability and Statistics course in the Korean high school curriculum immediately prior to the study and who expressed interest in statistical inquiry to participate in the study. The course introduced counting methods, probability (including conditional probability), probability distributions, and basic statistical inference. None of the nine students who agreed to participate in the study had prior experience analyzing data using technological tools. The learning sessions were facilitated by the first author. For five of the nine students, this was their first interaction with the first author. The students were familiar with one another and frequently assisted one another throughout the statistical investigation process.

3.2. TASK AND LEARNING ACTIVITIES

The task assigned to students in this study was to engage in a full cycle of statistical investigation using a secondary dataset titled "Physical Fitness Measurement Data" (Appendix A). Implemented over seven learning sessions, the task was guided by Lee et al.'s (2022) data investigation framework, which foregrounds core investigation practices including exploratory visualization. Reflecting Arnold's (2013) discussion of statistical investigations with secondary data—namely, that inquiry begins with familiarization with the dataset—the sequence started with examining the data collection context and variable meanings before moving to question formulation, analysis, and conclusion drawing.

The learning activities comprised seven 50-minute sessions conducted after school in mid-December, near the end of the second semester. Session 1 focused on exploring and understanding the dataset. Students watched a short video illustrating how big data can be used for prediction in different settings (e.g., baseball, U.S. elections) and then downloaded the dataset in CSV format from a website. They subsequently examined the metadata to identify who collected the data, how it was collected, and for what purpose (Buehring & Grando, 2023). Students were guided to interpret unfamiliar variables using supplementary materials and online searches. After this initial exploration, a preprocessed dataset was provided for subsequent analysis; variables with substantial missingness or requiring considerable domain-specific knowledge were removed from the dataset. In Session 2, students first shared topics and questions they wished to investigate during small-group discussions. As a whole class, they discussed the criteria that investigative questions should meet. The teacher also framed the purpose of the statistical investigation as "providing useful health-related information for students at our school," reflecting the criterion that students' questions should be worth investigating (Arnold, 2013).

In Session 3, students received a brief introduction to the use of CODAP as a tool for data analysis. The teacher introduced summary statistics (e.g., mean, standard deviation) and graphical displays (e.g., boxplots) for analyzing single variables as well as bivariate representations and measures (e.g., scatterplots, correlation coefficients, trend lines) for analyzing relationships between pairs of variables. For analyzing three or more variables, students were encouraged to coordinate visual tools such as boxplots and scatterplots—for example, by using side-by-side boxplots to compare the distribution of one variable across categories of another and scatterplots with points colored by a third variable (e.g., gender)—to explore multivariate relationships. They were also shown that, to represent a fourth variable, they could select the variable in the table and drag it onto the scatterplot, which activates yellow drop zones for assigning it (e.g., a secondary axis on the right-hand side).

In Session 4, students explored the dataset in CODAP by creating visualizations and computing descriptive statistics. The teacher asked students to draft their investigative questions and data analysis plans. In Session 5, the class revisited the criteria for investigative questions discussed in Session 2, and students reviewed whether their questions and analysis plans were aligned with the criteria. The activity report template was provided in this session, and students were asked to complete it. The teacher also briefly introduced data preprocessing to address dataset features such as missing values and outliers.

In Session 6, students explored and visualized the data, either individually or in groups, to address their investigative questions. Given that CODAP allows multiple variables to be represented within a single graph, students were guided to make use of its multivariable visualization features. In Session 7, students selected appropriate statistical measures (e.g., summary statistics, correlation coefficients, trend lines) and visualizations to answer their investigative questions and drew conclusions based on the resulting evidence.

3.3. DATASET AND TECHNOLOGICAL TOOLS

The dataset used in this study was a publicly available secondary dataset provided by the Korea Sports Promotion Foundation, which documented the procedures for data collection through its open data platform (https://www.bigdata-culture.kr/bigdata/user/data_market/detail.do?id=ace0aea7-5eee-

48b9-b616-637365d665c1). The data were collected from approximately 25,000 individuals who participated in National Fitness 100, a national physical fitness assessment program in which citizens participate to evaluate and improve their fitness. The dataset included over 60 physical-fitness variables, along with participant information spanning a wide age range, including both male and female participants (coded as gender in the dataset).

The variables included commonly recognized indicators such as age, gender, height, weight, body fat percentage, and waist circumference. Also included were physical fitness measures that may be less familiar to students and require additional interpretation, such as relative grip strength, the Illinois agility test, and hang time. The interquartile range for age showed that the majority of participants represented in the dataset were in their late teens to late thirties (Q1 = 16 years, Q3 = 39 years). Although the gender distribution was relatively balanced, there were slightly more male participants (12,847 males; 11,981 females). Below is a list of key variables referred to in this study and their descriptions:

- systolic blood pressure: maximum blood pressure when the heart contracts (mmHg)
- diastolic blood pressure: minimum blood pressure when the heart relaxes (mmHg)
- right-hand grip strength: grip strength measured with the right hand (kg)
- left-hand grip strength: grip strength measured with the left hand (kg)
- relative grip strength: grip strength of the stronger hand divided by body weight (%)
- BMI: body mass index, calculated as weight divided by the square of height (kg/m^2)
- body fat percentage: percentage of body weight that is fat tissue (%)
- hang time: time spent in the air during a jump (s)
- Illinois agility test: time taken to complete an agility test designed to measure agility and speed (s)

Students received the data files in two stages. First, they downloaded the raw dataset from the Korea Sports Promotion Foundation website, which supported independent exploration. Next, the teacher selected 26 of the 60 fitness variables to create a dimension-reduced dataset that was provided to students. (An example of the data is included in Appendix A.) Dimensionality reduction, a common preprocessing practice in data science (Franke et al., 2016), was used because many original variables were unfamiliar or difficult for students to interpret; variables with limited valid measurements were also excluded. The final dataset included 23 numerical and three categorical variables. This sequencing allowed students to first explore the raw data and then work with a reduced dataset in subsequent sessions.

The primary tool students used for data exploration and analysis throughout the learning activities was CODAP, with Excel used in the first two sessions to support initial data access and exploration. CODAP is well-suited for multivariate data (Biehler et al., 2024) and allows users to create visualizations via drag-and-drop, integrating visualization with analysis to support novice learners working with datasets (Sutherland & Ridgway, 2017). Because CODAP can display multiple representations simultaneously, it may support pattern identification and encourage active engagement in statistical investigation. However, CODAP did not always run smoothly when more than 5,000 observations were imported; therefore, students worked with random samples of 5,000 observations. Prior to introducing CODAP, Excel was used to help students view the dataset in tabular form and make initial sense of its structure.

3.4. DATA COLLECTION

Data collection occurred in two stages. First, after the seven-session sequence of learning activities was completed, students' activity reports and post-interviews were collected. The primary data source for this study was the activity reports; responses to post-interview questions were also included in the analysis to enrich the contextual understanding needed to interpret patterns of refinement in students' investigative questions. The activity report comprised multiple items; however, for this study, we collected only the items directly relevant to the research purpose: (1) focal investigative question, (2) helpful aspects and challenges in formulating the question, (3) examples of group cooperation during the investigation cycle, (4) data analysis plan, and (5) analysis results. Although nine students participated, some students submitted a joint activity report in pairs; therefore, seven activity reports were collected. For identification and analysis purposes, each report was assigned a code from A1 to A7.

The post-interviews were conducted via email, and responses to two questions, Questions 8 and 9, were collected as supplementary data to support interpretation. Because these two questions prompted students to reflect directly on (a) the criteria they applied when evaluating and revising their investigative questions and (b) their assumptions about the population when interpreting results, these questions were particularly useful for interpreting the refinement patterns observed in the activity reports. Question 8 asked, "We discussed what makes a good investigative question. Do you think your focal investigative question is a good one in light of that discussion? Please explain what aspects you think are strong and what aspects could be improved." Responses to this question were used to examine which aspects of students' focal investigative questions they foregrounded (or downplayed) when evaluating their questions, and the reasons they provided for those judgments. Question 9 asked, "Do you think the conclusion you drew can be generalized to the entire population of the country? Please explain your reasoning." Responses to this question were used to understand which population students had implicitly assumed in their investigative questions and why they may not have stated that population explicitly. Students responded to the email interview within one week of the final session, and eight participants submitted their responses (all except A4). To identify the source of each interview response, we used the following labeling convention. A1 and A2 refer to activity reports co-authored by two students. However, since the post-interviews were conducted individually, their responses were labeled as A1P1 and A1P2, and A2P1 and A2P2, respectively. In contrast, for students who worked individually (A3, A5, A6, and A7), the corresponding post-interview responses were labeled as A3P, A5P, A6P, and A7P.

Second, for analysis purposes, students' investigative questions were extracted from Items 1 and 4 of the activity reports (see Appendix B). The initiating investigative question (IIQ)—the question that motivated students to initiate the investigation and set its overall focus and direction—was extracted from Item 1. In Item 1, students were asked to articulate the investigative question that would guide their overall investigation, and IIQs were therefore typically presented as well-structured, sentence-level questions. The analysis-phase investigative question (AIQ) was derived from Item 4, which asked students to describe their plan for data analysis. AIQs emerged as students considered the requirements and constraints of the data analysis phase in relation to their IIQs, making them functionally distinct from IIQs. Because the purpose of Item 4 was to describe an analysis plan, the questions that appeared in this section were not always formulated as complete or well-structured interrogative sentences. We considered expressions such as "I want to analyze...," "I intend to check for correlations between…," and "I want to observe the correlation of…" in Item 4 to be AIQs when they could be formalized into investigative questions. For example, in A5's analysis plan, the student wrote the following:

> [After presenting three criteria for judging hypertension and hypotension status,] my judgment of hypertension and hypotension based on these criteria should be seen as only a possibility. However, since these criteria can somewhat indicate a relationship, I want to use them to examine it. [To do this,] I will use a scatterplot to show hypertension and hypotension status in relation to BMI and create a boxplot to examine how BMI differs by each status.

Therefore, we treated A5's response as two AIQs—"How does BMI differ by hypotension status?" and "How does BMI differ by hypertension status?"—and included both in our analysis.

Students were asked in Session 4 to formulate their investigative questions and data analysis plans, and an activity report template was provided in Session 5 to help them organize their work. The activity reports were collected at the end of the seven-session sequence. The IIQs and AIQs extracted from Items 1 and 4, which served as the main source of evidence for this study, may reflect not only students' own deliberation but also, to some extent, feedback from the teacher or peers during the sessions.

3.5. DATA ANALYSIS

Data analysis proceeded in three stages. First, the IIQs and AIQs were analyzed using the analytic framework developed from prior research (see Table 1). Except for variable-related components, each component of the analytic framework was applied once in coding each investigative question. In this study, multiple AIQs were often associated with a single IIQ, reflecting efforts to refine and operationalize the inquiry articulated in that IIQ during the course of data analysis. When coding the number of variables, we treated the set of AIQs associated with a single IIQ as a unified group. For example, if one IIQ was elaborated into three AIQs and five distinct variables were used across those three questions, the number of variables for that group was coded as five. The same group was used to identify the variables to be coded for variable type and clarity. However, because individual variables within a group may differ in type (e.g., categorical or numerical) and vary in their level of clarity, these two components were coded at the level of each variable identified within the group. If variables X, Y, and Z were identified in that group, each variable was examined and coded separately for type and clarity.

Table 1. Analytical framework for investigative questions during statistical investigation

Components		Analytical criteria
Variable number		The number of variables used in the question.
Variable type	Numerical	The variable used in the question can be expressed numerically.
	Categorical	The variable used in the question takes values that belong to a finite set of categories, which may or may not have an inherent order.
	Undetermined	The terms used in the question have the potential to be used as variables, but it is difficult to determine the type of variable.
Variable clarity	Clear	The variable(s) referred to in the question correspond directly to a dataset variable (or to a value calculated from dataset variables using an explicitly stated rule); the variable(s) can be analyzed without further specification.
	Somewhat clear	The variable(s) referred to in the question are related to variables in the dataset but require additional specification of how they will be defined or operationalized (e.g., selecting among candidate variables, stating how a variable is calculated or categorized).
	Unclear	The relationship between the variable(s) referred to in the question and the dataset variables is not clear, making the question difficult to address without substantial further operationalization.
Clarity of population	Clear	The population of interest is explicitly specified in the question with clearly identifiable boundaries.
	Somewhat clear	The population of interest exists but is ambiguously described in the question, requiring additional specification.
	Unclear	The population of interest cannot be identified from the question.
	Not relevant	The population of interest exists, but the scope of the question exceeds or is not structurally aligned with what the given data can represent, potentially suggesting inappropriate inference beyond the dataset.
Intent	Summary	The question is aimed at describing a specific attribute.
	Comparison	The question is aimed at comparing two or more variables.
	Relationship	The question is aimed at identifying the relationship between two or more variables.
	Prediction	The question is aimed at predicting a specific attribute.
Feasibility of drawing conclusions from the data	Feasible	The question is specific, uses variables in the dataset, and can be answered using the available sample without relying on external assumptions.
	Partially feasible	The question relates to the dataset but requires clarification, minor assumptions, or is limited by the available sample.
	Not feasible	The question cannot be answered due to missing variables, vague wording, or an insufficient sample size to support meaningful conclusions.
Global view of data	Feasible	The question allows drawing conclusions from an overall perspective on a group that contains variability rather than individual observations.
	Not feasible	The question focuses on one or a few specific individual observations, limiting the ability to draw conclusions about the overall group.

Note. Investigative questions were analyzed across five analytical components: variables, clarity of the population, intent, feasibility of drawing conclusions from the data, and global view of data. Within the variables component, three subcomponents were examined: number, type, and clarity.

Second, data from Items 2–5 of students' activity reports and from Questions 8 and 9 of the post-interview were analyzed using thematic analysis (Braun & Clarke, 2006) to examine how data visualizations appeared to support (or constrain) students' refinement of IIQs into AIQs, resulting in two themes related to representation and exploration. The data were read repeatedly to identify segments in which students described working with data visualizations and to explore how such segments could be connected to moments of question refinement; initial codes were generated accordingly. For example, the excerpt from A1P1's response to Question 8 presented in Episode 1 (see Section 4.2) was initially coded as "creating graphs using CODAP," "noticing that blood pressure was affected by age," and "using diastolic blood pressure instead." The first code contributed to the theme of representation because it referred to using visualization to make a relationship in the data visible. The latter two codes contributed to the theme of exploration because they reflected how the visualization helped reveal an age-related constraint in the data and informed a subsequent shift in the focus of analysis. Across the dataset, examples of codes associated with representation included "using regression lines" and "using coefficients of determination" (from Item 2 of A2; see Episode 3). Examples of codes associated with exploration included "questioning whether the findings could be generalized" and "noticing that the strength of correlations varied by age group" (from the response to Question 9 in A7P; see Episode 2). These and similar codes were compared and collated into the two themes, which were iteratively reviewed, defined, and named. Through this process, the final themes were organized around two broad functions of visualization in relation to question refinement: representation and exploration. Findings related to these themes are presented in Section 4.2 as episodes based on coded extracts from the dataset, each showing how data visualization appeared to support qualitative improvements across one or more of the five analytical components in the refinement of IIQs into AIQs.

Third, to enhance the credibility of the analysis, we conducted peer debriefing and collaborative review at two points. For the analytic-framework coding of IIQs and AIQs, the first author met online with two mathematics education researchers (both in-service teachers) and shared the full list of extracted questions, the coding framework, and examples of difficult cases. During this meeting, we discussed interpretive ambiguities (e.g., identifying variables and population referents) and resolved challenging coding decisions through negotiated agreement. For the thematic analysis of visualization's role in refinement, the first author developed the initial codes, candidate themes, and thematic definitions, which were then reviewed with the corresponding author. Instances of disagreement or insufficiently supported claims were revisited by returning to the original excerpts from activity reports and post-interviews, leading to refinement of code boundaries and theme labels. These revisions were incorporated into the final analysis.

4. RESULTS

This section is organized into two subsections to address the two research questions. The first subsection reports characteristics of the initiating investigative questions (IIQs) that students formulated in statistical investigations using secondary data and the associated analysis-phase investigative questions (AIQs). These characteristics are synthesized across five analytical components (see Appendix B for the full list of IIQs and AIQs). The second subsection provides three episodes based on students' activity reports and post-interviews.

4.1. IN A STATISTICAL INVESTIGATION USING SECONDARY DATA, HOW ARE STUDENTS' IIQS AND THEIR ASSOCIATED AIQS CHARACTERIZED ACROSS FIVE ANALYTICAL COMPONENTS?

To address our first research question, we analyzed a total of seven IIQs and 24 AIQs using the five analytical components presented in Table 1. The results for each component are reported in Table 2 (the variables component) and Table 3 (the remaining components). In the paragraphs that follow, selected examples are provided to highlight salient patterns and contrasts identified across the full set of IIQs and AIQs.

Table 2. Characteristics of investigative questions: Variables

		A1		A2		A3		A4		A5		A6		A7	
		IIQ	AIQ*	IIQ	AIQ*	IIQ	AIQ*	IIQ	AIQ*	IIQ	AIQ*	IIQ	AIQ*	IIQ	AIQ*
Variable number		1	6	0	4	1	6	4	5	3	3	0	5	4	4
Variable type	Numerical	1	6	–	4	1	6	4	5	1	1	–	5	4	4
	Categorical			–							2	–			
	Undetermined			–						2		–			
Variable clarity	Clear	1	6	–	4	1	6	4	5	1	3	–	5	3	3
	Somewhat clear			–						2		–		1	1
	Unclear			–								–			

Note 1. IIQ: Initiating investigative question; AIQ*: Analysis-phase investigative question group

Note 2. If the number of variables used in the question is zero, the type and clarity of variables are not classified and are marked as "–."

Table 3. Characteristics of investigative questions: Remaining analytical components

			A1					A2				A3						A4				A5			A6					A7			
			IIQ	AIQ				IIQ	AIQ			IIQ	AIQ					IIQ	AIQ			IIQ	AIQ		IIQ	AIQ				IIQ	AIQ		
				1	2	3	4		1	2	3		1	2	3	4	5		1	2	3		1	2		1	2	3	4		1	2	3
Clarity of population	Clear	●																												●	●	●	●
	Somewhat clear	◐	◐									◐	◐	◐	◐	◐	◐																
	Unclear	○		○	○	○	○		○	○	○							○	○	○	○	○	○	○		○	○	○	○				
	Not relevant	░						░																	░								
Intent	Summary	S	S																														
	Comparison	C																					C	C									
	Relationship	R		R	R	R	R		R	R	R		R	R	R	R	R	R	R	R	R	R				R	R	R	R		R	R	R
	Prediction	P						P				P													P					P			
Feasibility of drawing conclusions from the data	Feasible	●		●	●	●	●		●	●	●		●	●	●	●	●		●	●	●		●	●		●	●	●	●				
	Partially feasible	◐	◐															◐				◐								◐	◐	◐	◐
	Not feasible	○						○				○													○								
Global view of data	Feasible	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
	Not feasible	○																															

Note. Numbers 1-5 correspond to the AIQs listed in Appendix B. For example, A1's AIQ-1 refers to the first question in the list of AIQs derived from A1.

The characteristics of the variables included in the investigative questions were analyzed in terms of three components: number, type, and clarity. First, the analysis of the number of variables revealed that 13 variables were included in the IIQs, whereas the AIQ groups included 33 variables. On average, each IIQ contained 1.9 variables, and each AIQ group contained 4.7 variables. All seven AIQ groups included three or more variables, exceeding the number of variables identified in the corresponding IIQ. Second, in terms of variable type, the IIQs involved 11 numerical variables and two variables whose types could not be determined, whereas the AIQ groups involved 31 numerical variables and two categorical variables. In A5's IIQ, the two terms whose variable type could not be determined ("hypertension" and "hypotension") were represented in the AIQs as categorical variables—"hypotension status" and "hypertension status" (coded as TRUE/FALSE). A5 indicated that contextual knowledge about diagnostic criteria for hypertension and hypotension was applied to diastolic and systolic blood pressure to construct these two categorical variables. Third, in terms of variable clarity, 10 of the 13 variables in the seven IIQs were coded as clear, whereas 32 of the 33 variables in the AIQs were clearly specified. A small number of variables were classified as somewhat clear because, although they were related to dataset variables, they required further clarification before they could be operationalized. In some instances, this lack of clarity persisted from an IIQ into the associated AIQ group. For example, in both A7's IIQ and AIQs, the variable referred to as "grip strength" was not fully specified: the dataset contained three grip-strength variables (relative, right-hand, and left-hand grip strength), but the question did not specify which grip strength or combination of grip strengths was intended to represent the "grip strength" variable.

Clarity of the population was generally limited in both IIQs and AIQs. As shown in Table 3, A7 was the only activity report in which the population was stated clearly in the IIQ, and the analysis plan from which the AIQs were derived more clearly specified the population as 17–18-year-old high school students. In A3, the population could be identified in both the IIQ and the AIQs, but it was coded as somewhat unclear. The population was identified as "adolescents" in both IIQ and AIQs, but the question did not specify any age range for "adolescents." In A5, the IIQ was coded as unclear because the population was not stated, and the question focused on exploring the data to identify health-related facts rather than inferring health-related characteristics for a particular population. However, analysis of A5P indicated that, in response to Question 9, which asked whether the conclusion could be generalized to the entire population of the country, the student stated, "Yes. Since I analyzed 5,000 data points from a credible institution, I believe the objectivity is somewhat ensured." The student's "Yes" response suggests that the student regarded the population as the entire population of the country and was applying inference at that level. In A2, the IIQ was coded as presenting a population that was unrelated to the given dataset; the population was specified as "our school's students." In addition, the student's response to Question 9 in A2P1 indicated a lack of clarity about the scope to which the investigative questions could be generalized, as reflected in the following excerpt: "I think the second and third questions [AIQ-2 and AIQ-3] can be generalized. …" In supporting this claim, the student drew on prior contextual knowledge rather than drawing directly on the dataset. That is, the response indicated a tendency to rely more on prior beliefs or contextual knowledge than on the dataset in justifying the conclusion.

The IIQs reflected a range of intents, including summary, relationship, and prediction. Of these three intent types, prediction was the most frequent, followed by relationship (Table 3). An example of an IIQ categorized as having a prediction intent was A2's IIQ, which asked what students at their school should focus on to improve basic fitness. A4's IIQ reflected a relationship intent, as it asked about the correlation among grip strength, standing long jump, cross sit-ups, and body fat percentage. A1's IIQ, "What characteristics do people with high blood pressure generally have?" was the only question coded as having a summary intent, even when the corresponding AIQs were taken into account. When planning the data analysis, a single IIQ was often elaborated into multiple AIQs, and the AIQs elaborated from the same IIQ showed a consistent intent. Among the 24 AIQs, 22 had a relationship intent, and two had a comparison intent. For AIQs elaborated from a single IIQ and coded as reflecting a relationship intent, two patterns were observed in how relationships among the variables were specified in the AIQs. One pattern asked about the relationship between a focal variable X and another variable Y (e.g., A6), whereas the other included three or more variables and asked about relationships among them (e.g., A4). A5's AIQs were coded as reflecting a comparison intent because they asked how hypertension status (or hypotension status) varied according to BMI. Here, the variable hypertension status, unlike the numerical variable hypertension, had been reconstructed as a categorical variable consisting of two categories (TRUE/FALSE). That is, A5's AIQs were coded as questions comparing BMI between the group with hypertension (or hypotension) and the group without hypertension (or hypotension).

The feasibility of drawing conclusions from the data differed between the IIQs and the AIQs. In the IIQs, four were coded as partially feasible and three as not feasible (Table 3). By contrast, all AIQs were coded as feasible except the three AIQs in A7, which remained partially feasible. A3 provided an example of the contrast between an IIQ coded as not feasible and AIQs coded as feasible. The IIQ in A3—"Does relative grip strength have value as an indicator for assessing physical fitness and suggesting improvements in adolescents?"—signaled an interest in using relative grip strength as an indicator, but it did not specify how "physical fitness" would be operationalized using variables in the dataset. Without identifying which variables would serve as criteria for evaluating physical fitness, it was difficult to anticipate how statistically meaningful conclusions could be drawn from the data. In A3's Item 4 response, the student described physical fitness in terms of five elements (aerobic endurance, flexibility, body composition, muscular strength, and muscular endurance) and linked each element to variables in the dataset (e.g., body fat percentage and curl-ups). The student also specified an average grip strength variable computed by averaging right- and left-hand grip strength, which made the target variable more directly aligned with the variables available in the dataset. The A3 AIQs consisted of five questions examining relationships between relative grip strength and selected fitness-related variables (e.g., "What is the correlation between relative grip strength and body fat percentage in adolescents?"), and all of these questions were feasible for drawing conclusions from the data. In A7, both the IIQ and the AIQs were coded as partially feasible because the variable "grip strength" was used without clearly specifying which dataset variable(s) it referenced. Given that the dataset included multiple related measures—left-hand grip strength, right-hand grip strength, and relative grip strength—the term "grip strength" remained ambiguous, which limited the feasibility of drawing conclusions from the data.

Both IIQs and AIQs were coded as reflecting a global view of data (Table 3). For example, A5's IIQ focused on the relationship between BMI and hypertension/hypotension, which entailed examining patterns across multiple observations. Some investigative questions were phrased in a yes-or-no form but were treated as requiring interpretation across multiple observations. For instance, A3's IIQ asked whether relative grip strength has value as an indicator for assessing physical fitness, and its AIQs consisted of questions about relationships between relative grip strength and five variables (e.g., shuttle run, body fat percentage). On this basis, A3's IIQ and AIQs were coded as reflecting a global view because they involved evaluating relationships among variables across the dataset.

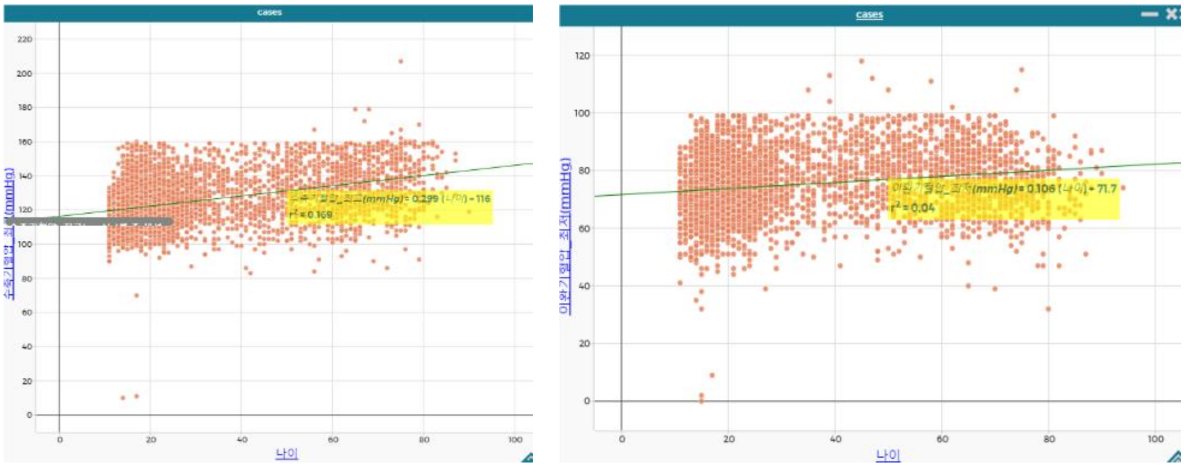
4.2. WITHIN THE SAME ACTIVITY, WHAT ROLE DOES DATA VISUALIZATION APPEAR TO PLAY IN STUDENTS' REFINEMENT OF IIQS INTO AIQS?

To address Research Question 2, we analyzed data from the activity reports (Items 2, 3, 4, and 5) and the post-interviews (Questions 8 and 9). We identified three episodes in the dataset in which data visualization appeared to support students' refinement of their IIQs into AIQs with respect to one or more of the five analytical components.

Episode 1 In A1, the IIQ used the general term "blood pressure," which was operationalized as the dataset variable "diastolic blood pressure" and explained more fully in A1P1. In response to Question 8, the student stated:

> Because diastolic and systolic blood pressure are influenced by age, I was not able to reach the desired conclusion when I created graphs using CODAP. (…) Since the dataset included a relatively large proportion of younger individuals, I used diastolic blood pressure [as a substitute for the blood pressure variable in the IIQ], which is less affected by age.

To illustrate what may have influenced the student's decision to focus on diastolic blood pressure, Item 5 of A1 presented CODAP visualizations of the relationships between age and systolic blood pressure (left panel) and between age and diastolic blood pressure (right panel), shown as scatterplots with least-squares regression lines (LSRL; see Figure 2) for the relationships. In addition, within A1's AIQ group, the broad term "characteristics" in the IIQ was elaborated into concrete variables, namely BMI, weight, and body fat percentage. The visual comparison in Figure 2 related to blood pressure suggested that diastolic blood pressure was a more appropriate way to operationalize the IIQ term "blood pressure" for the subsequent analysis. Compared with systolic blood pressure, diastolic blood pressure had a weaker relationship with age, a lower coefficient of determination, and a more diffuse pattern of points. This visual pattern suggests that specifying blood pressure as diastolic blood pressure supported the student in treating age as less likely to function as a confounding variable when examining relationships with other variables. In this way, the visualization was linked to a refinement in which the variable used in the IIQ became more clearly specified in the AIQ.



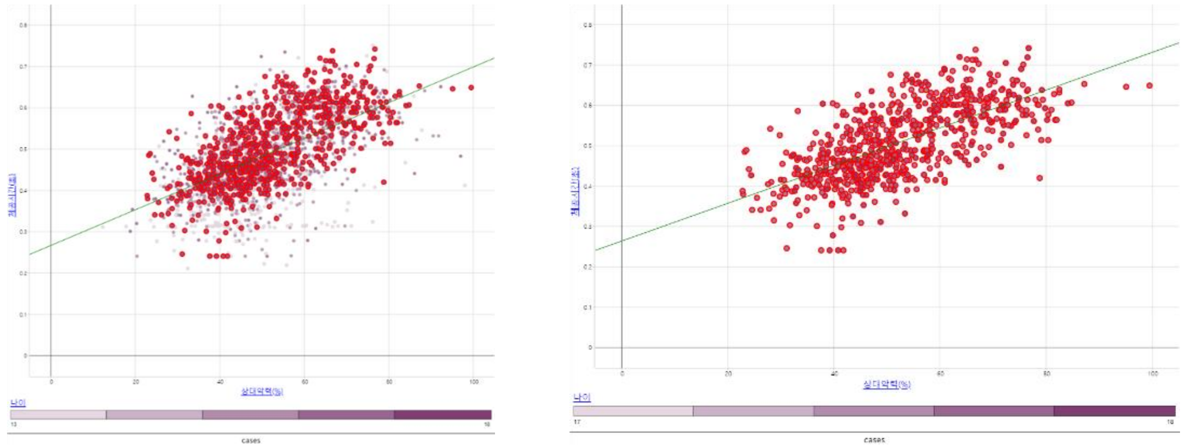
horizontal axis: age
vertical axis: systolic blood pressure
LSRL: systolic blood pressure in mmHg = 0.299(age) + 116
r^2=0.169

horizontal axis: age
vertical axis: diastolic blood pressure
LSRL: diastolic blood pressure in mmHg = 0.106(age) + 71.7
r^2=0.04

Figure 2. Data visualizations in A1

Episode 2 In Item 2 of A7, the student reported difficulty identifying variables that showed a meaningful correlation with grip strength and, using data visualizations, explained that the variables most strongly correlated with grip strength and the strength of those correlations differed across age groups (Figure 3). The scatterplot on the left in Figure 3 presents the relationship between relative grip strength and hang time, with the age range restricted to 13–18 years; the red points are data from 17–18-year-olds. The scatterplot on the right visualizes the same relationship using only the data for 17–18-year-olds. The student concluded that the correlation between the two variables appeared stronger for the 17–18-year-old subgroup compared to the broader 13–18-year-old group. Although the student's IIQ referred generally to "high school students," their AIQs referred more narrowly to "17–18-year-old high school students." The student deliberately extracted and analyzed only the data corresponding to this age group and then drew conclusions tailored to this subgroup. Furthermore, the student's response to Question 9 of A7P demonstrates a consistent recognition of the target population of 17–18-year-old high school students throughout the investigation:

> It is difficult [to generalize], because the correlation I investigated was based on 17–18-year-old high school students. The Korean population includes people of very diverse ages, and this leads to substantial physical differences. Therefore, it is not feasible to establish a single standard for physical training applicable to everyone. In fact, when I examined the correlations between relative grip strength and variables such as the Illinois agility test, standing long jump, and hang time for different age groups, the strength of the correlations varied.



horizontal axis: relative grip strength
vertical axis: hang time
color: red is data for selected 17–18-year-olds; grey is data for other ages (13–16) within the 13–18 filter

horizontal axis: relative grip strength
vertical axis: hang time
color: red is data for ages 17–18 (only)

Figure 3. Data visualizations in A7

The two visualizations of age-based subgroups seemed to support the student in exploring age-related variation and recognizing that the relationship under investigation was more interpretable for the 17–18-year-old subgroup than for the broader 13–18-year-old group. This visual contrast was linked to the student's redefining the broadly stated population in the IIQ and formulating AIQs that reflected a more clearly age-bounded population.

Episode 3 In A2, the IIQ asked: "What should our school's students improve to build basic fitness?" This IIQ was coded as having limited feasibility for drawing conclusions from the data because it was difficult to anticipate which dataset variables could be used to pursue the question in an analyzable way. In contrast, the AIQs were specified as questions about relationships between relative grip strength and variables such as body fat percentage, cross sit-ups, and standing long jump, and were coded as feasible for drawing conclusions from the data. In Item 2 of A2, the student documented that working with CODAP visualizations was connected to developing these investigative questions: "Using CODAP, I was able to examine the correlations between relative grip strength and body fat percentage, relative grip strength and cross sit-ups, and relative grip strength and standing long jump using least-squares regression lines and coefficients of determination."

The student also described a difficulty encountered during question formulation, noting that they had originally planned to examine the relationship between relative grip strength and weight but then realized that relative grip strength was computed by dividing grip strength by weight and multiplying by 100; the student reported that, as a result, the coefficient of determination was 0 in their analysis and presented a scatterplot of this relationship (see Figure 4). The process of visually exploring multiple variables was related to the student's ability to distinguish between variables to retain in the analysis and those to exclude, and, in this sense, to a partial refinement of the investigative questions.

horizontal axis: relative grip strength
vertical axis: weight

Figure 4. Data visualization in A2

5. DISCUSSION AND CONCLUSIONS

Prior research on investigative question refinement has mainly focused on instructional supports designed to help students improve the questions they pose at the outset of a statistical investigation. Although this work has recognized that such questions may be revised over the course of an investigation, relatively limited attention has been given to how the demands and constraints of the data analysis phase shape that revision. Addressing this gap, we distinguished between initiating investigative questions (IIQs) and analysis-phase investigative questions (AIQs) and examined how the broad questions students posed to initiate their investigations were refined into more analytically tractable forms. In doing so, this study highlights an aspect of question refinement that has remained empirically underexamined in classroom research: how demands and constraints that emerge during data analysis shape the reformulation of investigative questions. Our findings also indicate that, as IIQs were refined into AIQs, question quality improved in some respects, and that students' use of data visualization was associated with these partial improvements.

One of the most prominent changes in the refinement from IIQs to AIQs was that the questions became more clearly connected to the dataset variables and were restructured into forms better suited for drawing conclusions from the data. The average number of variables in the AIQs was more than twice that in the IIQs, which contrasts with prior research reporting that students rarely pose questions involving three or more variables, even when working with multivariate data (Arnold, 2013; Frischemeier & Biehler, 2018). Importantly, this increase in the number of variables did not simply reflect expansion, but was accompanied by greater specificity. For example, the everyday term "blood pressure" in A1 did not correspond directly to a single dataset variable, but in the AIQ it was operationalized as "diastolic blood pressure," thereby improving its alignment with the dataset. The AIQs were also better suited to drawing conclusions from the given dataset. For instance, in A3, a prediction-oriented IIQ about whether relative grip strength was a useful indicator of physical fitness was identified; in the AIQs, this question was elaborated into questions that broke "physical fitness" into five subcomponents and connected them to dataset variables. These AIQs, rather than maintaining the prediction intent as such, were restructured mainly into questions about bivariate relationships, and in that process became more closely aligned with the available dataset variables and more interpretable than the IIQ. This pattern highlights the importance of attending to how questions relate to the available dataset variables (Leavy & Frischemeier, 2022). In light of Laina and Wilkerson's (2016) argument that students may flexibly adjust their interpretations when making sense of complex visualizations containing multiple patterns, the present study suggests that, alongside shifts in interpretation, investigative questions themselves may also be reconfigured during the analysis process. These findings suggest that some initially vague or prediction-oriented IIQs, rather than being regarded simply as low-quality questions, may function as productive starting points that can be progressively clarified during the analysis process in terms of how they connect to the data and what kinds of conclusions can be drawn, thereby becoming more analyzable (Allmond & Makar, 2010).

Clarity of the population remained an area in need of improvement in both IIQs and AIQs, and this issue appeared more often in AIQs. This is consistent with Arnold's (2013) observation that students often have difficulty specifying the population in investigative questions. For example, in A5, although the target population was not stated in the AIQ, the student claimed in Question 9 of A5P that the conclusion could be generalized to a nationwide population. However, because the dataset consisted mainly of individuals in a particular age range, the scope of such generalization was limited. Although students had opportunities during the activity to attend to the target population, such as through exploring metadata and discussing criteria for good investigative questions, these difficulties persisted across both IIQs and AIQs. In addition, the task was framed as identifying health-related information relevant to students at their school, and this framing may have somewhat blurred the boundaries of the population of interest. These findings suggest the need to address more explicitly and repeatedly in classroom discourse the alignment among the goal of the task, the composition of the dataset, and the intended target population. More broadly, students often attempted informal statistical inference or tried to extend results beyond the sample when describing overall patterns in the data, but such accounts were not always aligned with a clearly specified population. This suggests that developing a global view of data in the context of large secondary datasets involves not only reading graphs or summary statistics, but also relating interpretations of patterns to judgments about the range over which generalization is warranted. This interpretation also connects with Bakker and Gravemeijer's (2004) argument that students' developing global view is intertwined with their thinking about sampling. Accordingly, teachers need to support students not only in describing distributions, but also in considering the scope and limits of their informal statistical inferences.

Our findings provide a classroom-based illustration of how students' use of a data visualization tool such as CODAP to generate, compare, filter, and adjust visualizations was associated with the refinement of investigative questions toward more analytically tractable forms (see Lee et al., 2022). In their representational role, the visualizations made distributions, variability, and relationships visible. This visibility was associated with students' identifying dataset variables that could operationalize everyday-language terms, as in Episode 1, and with setting aside variables that contributed little to drawing conclusions, as in Episode 3, after examining the relationships between focal variables and relative grip strength. In their exploratory role, CODAP's interactive features, particularly its filtering function, provided opportunities for students to revise the scope or focus of their questions. In Episode 2, the student plotted relative grip strength on the horizontal axis, hang time on the vertical axis, and age as a third variable, and then filtered the display to include only observations for 17–18-year-olds. This exploratory process was associated with reformulating the AIQ population as "17–18-year-old high school students," and in the post-interview, the student likewise limited the generalization of the conclusion to that subgroup. In this sense, the episode illustrates how exploratory work with visualizations may, in some cases, help students specify a target population more precisely.

This study is based on written documents produced by a small group of high school students. Because the participants showed particular interest in the data context and in statistical investigation activities, the setting may differ in some respects from more typical classroom contexts. Further research with more diverse groups of learners and instructional settings is therefore needed. In addition, the role of data visualization in supporting the global view of data component was not sufficiently captured here. This points to the need for further work on how visualizations may help learners coordinate descriptions of overall patterns with decisions about the warranted scope of generalization as they develop and refine investigative questions, as well as on how other digital resources, including generative AI tools, may shape this process as such tools become increasingly common in statistical investigations. As the use of secondary data continues to receive growing attention in statistics education, this study contributes to ongoing discussions about classroom integration by characterizing features of students' investigative questions and offering empirical insights into the role of data visualization in their refinement.

REFERENCES

Allmond, S., & Makar, K. (2010). Developing primary students' ability to pose questions in statistical investigations. In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the eighth International Conference on Teaching Statistics (ICOTS8, July 2010), Ljubljana, Slovenia.* International Statistical Institute. https://icots.info/icots/8/cd/pdfs/invited/ICOTS8_8A1_ALLMOND.pdf

Arnold, P. M. (2013). *Statistical investigative questions: An enquiry into posing and answering investigative questions from existing data* [Doctoral dissertation, University of Auckland]. https://researchspace.auckland.ac.nz/handle/2292/21305

Arnold, P., & Franklin, C. (2021). What makes a good statistical question? *Journal of Statistics and Data Science Education, 29*(1), 122–130. https://doi.org/10.1080/26939169.2021.1877582

Bakker, A., & Gravemeijer, K. P. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 147–168). Springer. https://doi.org/10.1007/1-4020-2278-6

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II)*. American Statistical Association; National Council of Teachers of Mathematics. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEIIPreK-12_Full.pdf

Bargagliotti, A., & Gould, R. (2022). Secondary data in the secondary data science and statistics classroom. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering and educating today's learners in statistics. Proceedings of the eleventh International Conference on Teaching Statistics (ICOTS11, 2022), Rosario, Argentina.* International Association for Statistical Education. https://icots.info/icots/11/proceedings/pdfs/ICOTS11_127_GOULD.pdf?1669865522

Ben-Zvi, D., & Arcavi, A. (2001). Junior high school students' construction of global views of data and data representations. *Educational Studies in Mathematics, 45* (1–3), 35–65. https://doi.org/10.1023/A:1013809201228

Biehler, R., Frischemeier, D., Gould, R., & Pfannkuch, M. (2024). Impacts of digitalization on content and goals of statistics education. In B. Pepin, G. Gueudet, & J. Choppin (Eds.), *Handbook of digital resources in mathematics education* (pp. 547–583). Springer. https://doi.org/10.1007/978-3-031-45667-1_20

Bolch, C., & Crippen, K. (2022). Data scientists' epistemic thinking for creating and interpreting visualizations. *Statistics Education Research Journal, 21*(2), Article 11. https://doi.org/10.52041/serj.v21i2.21

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Buehring, R. S., & Grando, R. C. (2023). Reading and writing the world with children: Statistical thinking and multivariate data. *Statistics Education Research Journal, 22*(2), Article 6. https://doi.org/10.52041/serj.v22i2.446

Cairo, A. (2012). *The functional art: An introduction to information graphics and visualization*. New Riders.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: Using vision to think*. Morgan Kaufmann.

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 375–396). Springer. https://doi.org/10.1007/1-4020-2278-6

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801–823. https://doi.org/10.1080/00029890.1997.11990723

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16*(1), 44–49. https://doi.org/10.52041/serj.v16i1.213

Fielding, J., Makar, K., & Ben-Zvi, D. (2025). Developing students' reasoning with data and data-ing. *ZDM Mathematics Education, 57*(1), 1–18. https://doi.org/10.1007/s11858-025-01671-6

Franke, B., Plante, J. F., Roscher, R., Lee, E. S. A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M. M., Grosse, R., Hendricks, D., & Reid, N. (2016). Statistical inference, learning and models in big data. *International Statistical Review, 84*(3), 371–389. https://doi.org/10.1111/insr.12176

Frischemeier, D., & Biehler, R. (2018). Stepwise development of statistical literacy and thinking in a statistics course for elementary preservice teachers. In T. Dooley & G. Gueudet (Eds.), *Proceedings of the 10th Congress of the European Society for Research in Mathematics Education* (pp. 756–763). DCU Institute of Education; ERME. https://hal.archives-ouvertes.fr/hal-01927856

Frischemeier, D., & Leavy, A. (2020). Improving the quality of statistical questions posed for group comparison situations. *Teaching Statistics, 42*(2), 58–65. https://doi.org/10.1111/test.12222

Graham, A. (2006). *Developing thinking in statistics*. Paul Chapman.

Hall, J. (2011). Engaging teachers and students with real data: Benefits and challenges. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics: Challenges for teaching and teacher education* (pp. 335–346). Springer. https://doi.org/10.1007/978-94-007-1131-0_32

Higgins, T., Mokros, J., Rubin, A., & Sagrans, J. (2023). Students' approaches to exploring relationships between categorical variables. *Teaching Statistics, 45*(S1), S52–S66. https://doi.org/10.1111/test.12331

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kazak, S., Fujita, T., & Turmo, M. P. (2023). Students' informal statistical inferences through data modeling with a large multivariate dataset. *Mathematical Thinking and Learning, 25*(1), 23–43. https://doi.org/10.1080/10986065.2021.1922857

Kim, A. Y., Ismay, C., & Chunn, J. (2018). The fivethirtyeight R package: "Tame data" principles for introductory statistics and data science courses. *Technology Innovations in Statistics Education, 11*(1). https://doi.org/10.5070/T5111035892

Korea Sports Promotion Foundation. (n.d.). *Physical fitness measurement data*. Bigdata-Culture. https://www.bigdata-culture.kr/bigdata/user/data_market/detail.do?id=ace0aea7-5eee-48b9-b616-637365d665c1

Laina, V., & Wilkerson, M. H. (2016). Distributions, trends, and contradictions: A case study in sensemaking with interactive data visualizations. In C. K. Looi, J. L. Polman, U. Cress, & P. Reimann (Eds.), *Transforming learning, empowering learners: The International Conference of the Learning Sciences (ICLS) 2016* (Vol. 2, pp. 934–937). International Society of the Learning Sciences. https://repository.isls.org/bitstream/1/347/1/140.pdf

Leavy, A., & Frischemeier, D. (2022). Developing the statistical problem posing and problem refining skills of prospective teachers. *Statistics Education Research Journal, 21*(1), Article 10. https://doi.org/10.52041/serj.v21i1.226

Lee, H., Mojica, G., Thrasher, E., & Baumgartner, P. (2022). Investigating data like a data scientist: Key practices and processes. *Statistics Education Research Journal, 21*(2), Article 3. https://doi.org/10.52041/serj.v21i2.41

Lee, V. R., & Wilkerson, M. (2018). *Data use by middle and secondary students in the digital age: A status report and future prospects.* National Academies of Sciences, Engineering, and Medicine. https://digitalcommons.usu.edu/itls_facpub/634/

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

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation* (4th ed.). Jossey-Bass.

Podworny, S., Fleischer, Y., Stroop, D., & Biehler, R. (2022). An example of rich, real and multivariate survey data for use in school. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (Eds.), *Proceedings of the twelfth Congress of the European Society for Research in Mathematics Education (CERME12)* (pp. 940–947). ERME; Free University of Bozen-Bolzano. https://hal.science/CERME12/hal-03751842v1

Sutherland, S., & Ridgway, J. (2017). Interactive visualisations and statistical literacy. *Statistics Education Research Journal, 16*(1), 26–30. https://doi.org/10.52041/serj.v16i1.210

The Concord Consortium. (2014). *Common Online Data Analysis Platform (CODAP)* [Computer software]. https://codap.concord.org/

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.
Watson, J. M., & English, L. D. (2017). Statistical problem posing, problem refining, and further reflection in grade 6. *Canadian Journal of Science, Mathematics and Technology Education, 17*(4), 347–365. https://doi.org/10.1080/14926156.2017.1380867
Watson, J., & Fitzallen, N. (2015). Statistical software and mathematics education: Affordances for learning. In L. D. English & D. Kirshner (Eds.), *Handbook of international research in mathematics education* (3rd ed., pp. 563–594). Routledge. https://doi.org/10.4324/9780203448946-29
Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x
Wilke, C. O. (2019). *Fundamentals of data visualization: A primer on making informative and compelling figures*. O'Reilly Media.
Zapata-Cardona, L. (2025). Public engagement of underserved students with open civic data. *ZDM Mathematics Education, 57*(1), 19–30. https://doi.org/10.1007/s11858-024-01641-4

HYEJIN JUN
Seoul National University
1 Gwanak-ro
Gwanak-gu
Seoul, Republic of Korea

Appendix A: Example from the dataset

Measure place	Age	Grade	Gender	Height (cm)	Weight (kg)	Body fat(%)	Waist circumfe rence (CM)	Diastolic blood pressure (mmHg)	Systolic blood pressure (mmHg)	Grip strength _left (kg)	Grip strength _right (kg)	Sit-ups	Repeated jumps	Seated sit-ups	Illinois agility test	Hang time	Coordina tion time	Coordina tion errors	Coordina tion calculate d value	BMI (kg/m^2)	Cross sit-ups	Standing long jump	Figure-eight walking	Relative grip strength	Whole-body reaction time
Jeju	30	participate	F	167.6	52.56	16	68.2	85	128	19.3	20			9.1						18.7	32	137		38	
Seodaem	24	3	F	171.4	64.8	25.1	68.2	68	103	29.7	30.9			10.4						22.1	36	163		47.7	0.38
Siheung	68	3	M	171.8	103.7	34.9	116.8	84	152	52.9	55.7			2.2						35.1			19.53	53.7	
Ansan	74	3	M	166.5	78.7	34.8	101	69	149	34.8	41.6			3						28.4			22.8	52.9	
Hwaseong	25	3	M	174.6	77.9	19.8		93	141	39.5	41.2			11.4						25.6	41	222		52.9	0.291
Namgu(B	20	1	M	176.3	69.1	16	77	79	127	47.5	46			18.7						22.2	56	257		68.7	0.328
Saha	18	2	M	174.3	59.9	18.6	77	75	126	37.8	39.1		53	16.7	18.38	0.626	35.841	6	36.117	19.7				65.3	
Muan	71	3	M	164.3	66.7	24.4	87	72	133	34.3	27.2			10.6						24.7			25.44	51.4	
Seongdor	23	participate	F	164.7	51.6	25.3		57	110	23.9	26.3			27.6						19	20			51	0.368
Seongdor	60	participate	M	168	62.3	19.2		83	123	36	37.2			-2.5						22.1	44	211		59.7	
Jungu(Se	27	participate	M	170.3	69.4	27.1		82	133	40.7	44.2			5.3						23.9	29	194		63.7	0.358
KSPO Son	27	3	F	169.6	50.4	23		73	131	18.3	23.5			12.9						17.5	31			46.6	0.334
Mapo	60	2	F	144	56.8	38.7		81	143	21.7	24.8			20.9						27.4	25	132		43.7	0.301
Choongju	18	3	M	168.8	66.8	17.6	75	49	124	44.1	40		48	6.1	20.75	0.533	53.595	7	53.917	23.4				66	
Seogu(gw	72	3	F	154	62	39.6	89	69	121	20.6	19.1			23.5						26.1			21.82	33.2	
Bukgu(gw	40	participate	M	174.7	74.8	24.6		88	123	38.5	38.5			7.4						24.5	25			51.5	0.526
Ansan	50	participate	F	160.1	93.4	48.5	113	108	170	23.1	29.6			-0.1						36.4	0	137		31.7	0.518
KSPO Son	29	3	F	154.5	55.5	37.2		83	135	21.8	22.9			13.7						23.3	27			41.3	0.41
Hwaseong	74	3	M	164.8	60.5	21.4		70	146	15.5	28.3			12.7						22.3			23.85	46.8	
Seogu(gw	65	participate	F	158.5	50.7	22.4	73	91	124	21.6	24.9			22						20.2			20.98	49.1	
Choongju	25	participate	M	169.6	57.4	16.1	72	64	111	40.3	40.5			2.1						20	40	213		70.6	0.353
Gumi	33	participate	M	170.7	73.1	24.6	91.8	87	136	55.2	52.8			-0.3						25.1	44	209		75.5	0.295
Sacheon	29	participate	M	171.4	99.4	36.5	118.3	96	156	48.7	52			8.7						33.8	44			52.3	0.378
Yangpyeo	62	participate	M	155.9	53.5	20.7	68	71	134	21.1	26.2			-7.7						22	15			49	0.425
Jeju	25	participate	M	186.9	110.52	26.9	111.4	80	151	52.2	60.1			10.7						31.6	41	211		54.3	
Samchuck	74	2	F	157	59.5	33.4		89	156	24.1	27.6			20.8						24.1			25.92	46.4	
Samchuck	72	participate	F	149	58.1	41.7		96	154	20.5	23.8			-15.5						26.2			26.61	41	
Samchuck	69	participate	F	154.5	60.7	40.9		85	155	16.1	17			14						25.4			28.49	28	
Naju	64	participate	M	166.5	70	23.2	81.8	76	138	40.9	46.3			-12						25.3	23	170		66.1	
Seogu(gw	56	participate	F	157.8	68.4	39.1	90	93	140	22.2	26			23						27.5	24	135		38	0.328
Donggu(B	49	participate	M	166.8	77.7	22.9	91	82	135	45.1	51.8			-5.7						27.9	33			66.7	0.307
Gunsan	44	2	M	178.3	65.7	10.9		93	142	37.8	44.8			13.6						20.7	47	223		68.2	0.31

Note. Blank cells indicate that the measure was not recorded/available for that participant (i.e., missing data), not a value of zero. See Section 3.3 for details about the dataset and variable descriptions.

Appendix B: Investigative questions formulated by students

	Initiating investigative question		Analysis-phase investigative question
A1	What characteristics do people with high blood pressure generally have?	1)	What is the relationship between age and systolic blood pressure?
		2)	What is the relationship between BMI and diastolic blood pressure?
		3)	What is the relationship between weight and diastolic blood pressure?
		4)	What is the relationship between body fat percentage and diastolic blood pressure?
A2	What should our school's students improve to build basic fitness?	1)	What is the correlation between relative grip strength and body fat percentage?
		2)	What is the correlation between relative grip strength and cross sit-ups?
		3)	What is the correlation between relative grip strength and standing long jump?
A3	Does relative grip strength have value as an indicator for assessing physical fitness and suggesting improvements in adolescents?	1)	What is the correlation between relative grip strength and the shuttle run in adolescents?
		2)	What is the correlation between relative grip strength and the sit-and-reach test in adolescents?
		3)	What is the correlation between relative grip strength and body fat percentage in adolescents?
		4)	What is the correlation between relative grip strength and average grip strength in adolescents?
		5)	What is the correlation between relative grip strength and curl-ups in adolescents?
A4	What is the correlation among grip strength, standing long jump, cross situps, and body fat percentage, and is this correlation significant? If so, how can this information be utilized?	1)	What is the correlation among right-hand grip strength, standing long jump, and body fat percentage?
		2)	What is the correlation among right-hand grip strength, cross sit-ups, and body fat percentage?
		3)	What is the correlation among right-hand grip strength, relative grip strength, and body fat percentage?
A5	What is the correlation between BMI and hypertension, and between BMI and hypotension, respectively?	1)	How does BMI differ by hypotension status?
		2)	How does BMI differ by hypertension status?
A6	What can the students in our class do to jump higher and longer during basketball games?	1)	What is the correlation between body fat percentage and hang time?
		2)	What is the correlation between right-hand grip strength and hang time?
		3)	What is the correlation between standing long jump and hang time?
		4)	What is the correlation between body weight and hang time?
A7	Can appropriate exercise targets (agility, leg strength) be suggested based on grip strength for high school students (e.g., standing long jump, hang time, Illinois agility test)?	1)	What is the correlation between 17–18-year-old high school students' grip strength and their Illinois agility test results?
		2)	What is the correlation between 17–18-year-old high school students' grip strength and their standing long jump performance?
		3)	What is the correlation between 17–18-year-old high school students' grip strength and their hang time?

Note. In their activity reports, A1 and A4 specified the dataset variables used to operationalize an IIQ variable: A1 mapped "blood pressure" to diastolic blood pressure, and A4 mapped "grip strength" to right-hand grip strength.