UNIVERSITY OF CANBERRA

Indonesian Year 9 and Year 12 Students' Statistical Literacy: Levels, Challenges and Understandings

Achmad Badrun Kurnia

A Thesis submitted in fulfilment for the degree of
Doctor of Philosophy in Mathematics Education

Faculty of Education

University of Canberra, Australia

June 2024

Abstract

In the information age, heightened during the COVID-19 pandemic, the capability to interpret and critically engage with data-driven information become an essential skill. This study aimed to address this crucial issue by focusing on the statistical literacy (SL) of Indonesian students, a population often underrepresented in research. Notably, the Programme for International Student Assessment (PISA) 2003 and 2012 tests have shown that Indonesia was among the underperforming developing countries in the uncertainty and data subscale. This study is timely given that there was no recent PISA data available when this study began, and the most recent PISA data—which similarly showed Indonesian students' underperformance—was only recently released in 2023. In addition, there is a concern that Indonesian students' underachievement in SL has not improved substantially in their final years of formal education.

The study introduced a novel framework for SL assessment that is innovative in its comprehensive approach. This framework aimed to gauge not only the SL levels of Indonesian Year 9 and Year 12 students, but also to identify the specific challenges they faced and understandings they demonstrated. SL was determined through four complex response skills—interpreting, communicating, evaluating and decision-making—all of which were founded on three interrelated knowledge components: text and context, representation and statistical-mathematical knowledge. The framework incorporated a six-level hierarchy for each component. The lower three levels—idiosyncratic, informal and inconsistent—served to highlight the challenges students encountered, while the upper three levels—consistent non-critical, critical and critical mathematical—shed light on students' understandings.

To ensure a robust and diversified sample, the study adopted a stratified purposive and convenience sampling strategy, 96 students were drawn from 16 schools. The stratification included the students’ grade levels, gender, school type, school status and city of origin. The study was a cross-sectional study and employed an explanatory sequential design, starting with a quantitative component and subsequently delving into qualitative component. A test was administered, and a follow-up interview was undertaken to clarify students’ thought processes during the test. In the quantitative component, analyses included double coding of students' written responses, descriptive statistics and the application of the Mann-Whitney U test for non-parametric data. For the qualitative component, the study employed the four stages of the Constant Comparison Method (CCM) to gain nuanced insights into the students' written responses and subsequent interviews.

The results revealed that Year 12 students displayed statistically significant higher levels of SL and skills, except in interpreting. Furthermore, the study found no significant gender-based or other demographic-based differences in SL and skill levels. Qualitatively, the challenges and understandings in the four skill areas were closely linked to the students' appreciation of the three foundational knowledge components. The level of sophistication in one component appeared to influence the level of sophistication in the others. Most students in the lower group encountered challenges with contextual-graphical interrelationships, while students’ critical understandings of the context improved their ability to comprehend data presented in graphs and tables, and vice versa.

In summary, this study contributed a groundbreaking framework for the assessment of SL, one that has the potential to be broadly applied by educators for both evaluative and pedagogical purposes. The framework filled a significant research gap and had far-reaching implications for educational strategies and curriculum development aimed at promoting SL.

Certificate of Authorship of Thesis

Except where clearly acknowledged in footnotes, quotations and bibliography, I certify that I am the sole author of the thesis submitted today entitled:

Indonesian High School Students' Statistical Literacy: Levels, Challenges and Understandings

I further certify that to the best of my knowledge the thesis contains no material previously published or written by another person except where due reference is made in the text of the thesis.

The material in the thesis has not been the bases of an award of my other degree or diploma except where due reference is made in the text of the thesis.

The thesis complies with University of Canberra requirements for a thesis as set out in the *Examination of Higher Degree by Research Theses Policy:* https://www.canberra.edu.au/research/current-research-students/hdr-policy-and-procedures

Candidate's Signature   Date: 03 / Nov / 2023

T. Lowrie

Primary Supervisor's Signature   Date: 03 / Nov / 2023

Disclaimer

The research materials and data in this study, including test items, interview scripts, students' written responses and students' interview data, have been translated from Bahasa Indonesia into English.

Acknowledgments

Many people may love mathematics but hate statistics because it is too complex and boring, and I was once among this group of people. Twenty years ago, when I was still an undergraduate student studying mathematics education, statistics was a subject that I disliked the most. For me, it was too abstract, and I could not see any benefit from studying it except for a thesis at the end of my bachelor’s program. Convinced of its uselessness, I chose not to undertake a quantitative study for my bachelor thesis. However, eight years later, during my master’s program, life brought me closer to statistics and my passion for it grew as I realised its many uses. This growing love motivated me to pursue this PhD focusing on statistical literacy.

This PhD journey was one of the best experiences of my life; yet the process itself was not as smooth and easy as it seemed. Many things happened during this period of time, including the COVID-19 pandemic which provided a clear example of the importance of statistics in our daily lives. The COVID-19 outbreak nearly changed how my PhD journey concluded. I took a study leave that lasted more than a year to spend time with my family because of this outbreak. With all these happenings, this PhD thesis would not have been completed without the help of many people. To all of them, I would like to express my sincere thanks.

First of all, I would like to express my sincere gratitude to the Australian Department of Foreign Affairs and Trade (DFAT) and my supervisory panel. I am greatly honoured to be selected as one of the Australia Awards recipients to pursue a PhD program at the University of Canberra (UC). I am even more honoured to have three incredibly supportive supervisors. My primary supervisor, Centenary Professor Thomas Lowrie, always encouraged, guided and motivated me to work diligently, independently and responsibly for more than five years of

my study. His criticism and suggestions on my writing have shaped my intellectual abilities to a higher level every time we met in a supervision. My secondary supervisor, Dr. Sitti Maesuri Patahuddin, was an amazing researcher who required me a great deal of effort, an extensive amount of time and patience to meet her quality standards. When I felt confident with my writing, she would often ask me easily with ‘so what?’ question which was never easy to answer. My advisor, Dr Tracy Logan, was the one who accompanied me when I was confused and helpless. She indirectly taught me how to be critical without offending feelings and how to admit mistakes without feeling ashamed.

Second, I am immensely thankful to all colleagues at UC who supported me academically and non-academically. I thank the staffs of the Faculty of Education: Eleni Petraki, Deborah Pino Pasternak, Elke Stracke and Kylie Reece, who have been helpful in all study-related matters. I thank Timea Hrivik, a Student Contact Officer at UC, for her continuous support throughout my doctoral program. She always checked on me and helped me to plan anything related to scholarship. She and her team were the best in assisting me as an Australia Awards recipient. I also thank all team members at the STEM Education Research Centre, University of Canberra (SERC-UC) for their endless friendship during my doctoral program. In particular, I thank Annabel Fagence who was always able to find the best day and time for me to meet Tom. Special thanks go to the IndoPhD students at UC. I was so blessed with their constant support throughout this PhD. We shared our journey and milestones to strengthen each other, held weekend writing to speed up our progress, stayed together until midnight or even early morning in building 6 to be more productive on a quiet evening. All these memories will remain forever.

Third, I must express my gratitude to many people involved in data collections, analyses and editing. I thank school principals and mathematics teachers for their permission and students for their participation in this study. I am deeply indebted to the two research assistants who helped me analyse the data during approximately one year of online discussion, between me in Canberra and them in Indonesia. I also thank editors from Capstone Editing that provided copyediting and proofreading services according to the guidelines laid out in the university-endorsed national ‘Guidelines for Editing Research Theses’. Without their helps, this study would not have been accomplished.

Fourth, I would like to thank many people who supported me during my stay in Canberra, helping me to have a balanced life, between academic and social life. I thank Pak Marpudin, Pak Imam, Bu Ina, Bu Dina, Bu Nana and Pak Ihsan from the Australia-Indonesia Muslim Foundation in ACT (AIMFACT) who welcomed me warmly as a family and inspired me with all their social activities. I am grateful to colleagues from the Indonesian-Muslim organisation at ANU (*Khataman*) and at UC (UCKUM) and from the Muslim Student and Staff Association at UC (MSSAUC) who gave me a feeling of togetherness. I thank my fellow recipients of the Australia Awards who always motivated me, and it was so touching that you all came to celebrate my farewell. I also thank all the staffs at the Indonesian Embassy, thanks for your warm welcome.

Finally, to my family, I can never thank you enough for all the support, patience, understanding and trust you have given me. I am sure my late parents always prayed for my success, especially my mother who passed away in 2021 due to the COVID-19 viruses. I still remember the moment I witnessed your last breath alone in the hospital. I also thank my sisters and brothers who were always ready to help my wife and children when they needed a helping hand. I remember when you stayed in the hospital taking care of my children who were hospitalised because of typhus and *Dengue Hemorrhagic Fever*. I cannot imagine how I would get through such stressful situations, without your support, while I was in Canberra. Most importantly, I must convey my deepest thanks to my wife and my two children. Their love, support, patience, understanding, and trust motivated me to finish this PhD as quickly as

I can. Even though it took me longer than the average time, they still put their trust in me. My wife, Sarifah Mahdalena, was a strong woman and whole-heartedly committed to our decision to have long distance relationship for many years. Even if it was very difficult, she took care of our children by herself and gave them comfortable feelings as if I was with them. My daughter, Rifdha Shagufta Kurnia, always remembered the day I left her for Australia, when she was about to enter year 1. She also recalled my promise to get my PhD by the time she entered year 4, although she is now already in year 6. My son, Abbas Atha'illah Kurnia, was around two years old when I left him. He knew that his father was in Australia, but he did not know how far Australia is from Indonesia. When I was next to him, he often asked me 'How long will it take to finish your homework?', I had no better answer than 'soon and just pray for me'. Now, I have a better answer for him: 'I'm done, thank you.'

On the occasion of our 15th wedding anniversary, I dedicate this doctoral thesis to my wife. We were married on 13th June 2009, and on 13th June 2024, I received an email confirming my study completion.

Publication and Presentations from This Thesis

Publication

Kurnia, A. B., Lowrie, T. & Patahuddin, S. M. (2023). The development of high school students' statistical literacy across grade levels. *Mathematics Education Research Journal*, 1-29. https://doi.org/10.1007/s13394-023-00449-x

Short presentations

Kurnia, A. B. (2023). Cognitive interviews for item development. *Seminar Nasional: Best practices untuk pengajaran kreatif*. Fakultas Tarbiyah dan Ilmu Keguruan, Institut Agama Islam Negeri Langsa (IAIN Langsa), Indonesia (Oral presentation).

Kurnia, A. B. (2022). Can You Explain the 'Mean' Do You Mean?. *Faculty of Education Research Conference 2022 University of Canberra*. Australia (Oral presentation).

Kurnia, A. B. (2021). Lessons from The Students' Responses to Data-Based Items. *A Webinar on Statistical Thinking: Statistical Thinking Application in Education*. Southeast Asian Ministers of Education Organization, The Regional Centre for Education in Science and Mathematics (SEAMEO RECSAM), Malaysia (Oral presentation).

Kurnia, A. B. (2021). Examining The Statistical Literacy of High School Students. *Collaborative Research Webinar in Mathematics Education: Universitas Negeri Malang and University of Canberra*, Indonesia (Oral presentation).

Kurnia, A. B. & Patahuddin, S. M. (2018). Examining Statistical Literacy Skills Within High-Stakes Assessment. *The 41st annual conference of the Mathematics Education Research Group of Australasia (MERGA 41)*, New Zealand (Oral presentation).

Contents

List of Tables

List of Figures

List of Abbreviations and Terminologies

Abbreviations

5L	Let’s observe, let’s ask, let’s reason, let’s find information and let’s communicate
CCM	Constant Comparison Method
GAISE	Guidelines for Assessment and Instruction in Statistics Education
IELTS	International English Language Testing System
IndoMS	Indonesia Mathematical Society
K13	*Kurikulum* 2013 (Curriculum 2013)
LOCUS	Levels of Conceptual Understanding in Statistics
MGMP	*Musyawarah Guru Mata Pelajaran* (High School Subject Teaching Working Group)
MoEC-RT	Ministry of Education, Culture, Research and Technology
MoRA	Ministry of Religious Affairs
PISA	Programme for International Student Assessment
QIR	Qualitative Item Review
SL	Statistical Literacy
TIMSS	Trends in International Mathematics and Science Study
UKG	*Uji Kompetensi Guru* (Teachers’ Competency Test)
UN	*Ujian Nasional* (Indonesian National Examination)

Terminologies

Year 9 UN	UN for Indonesian Year 9 students
Year 12 UN	UN for Indonesian Year 12 students

High school	Years 7 to 12
Junior high school	Years 7 to 9
Senior high school	Years 10 to 12
Four SL skills	Interpreting, communicating, evaluating and decision-making
Three SL components	Text and context, representation and statistical-mathematical knowledge
Six hierarchical levels	The levels for students’ SL ranging from L1 (*idiosyncratic*), L2 (*informal*), L3 (*inconsistent*), L4 (*consistent non-critical*), L5 (*critical*) and L6 (*critical mathematical*)
Lower group level	L1 to L3
Upper group level	L4 to L6
Descriptor	The characteristics for each of the six hierarchical levels
Component-based descriptor	The descriptors that are based on the three SL components
Component-based item descriptor	The component-based descriptor for certain item

Chapter 1: Introduction

1.1 Rationale for the Study

The rapid dissemination of data-based information through technology requires students to be literate in statistics. Being statistically literate means being able to respond critically to information involving statistics, and many students and citizens find this challenging (Muñiz-Rodríguez et al., 2020; Shields, 2005). Moreover, technologies and the COVID-19 pandemic have increased the amount of data-based information to which students must respond (e.g., Büscher, 2022a; da Silva et al., 2021; Franklin, 2021; Franklin & Bargagliotti, 2020; Gonda et al., 2022; Sharma et al., 2011; Sharma, 2013b; Suarez-Alvarez, 2021; Watson & Callingham, 2020; West & Bergstrom, 2020). For instance, students are overwhelmed by predictions and claims based on the numbers of COVID-19 cases, deaths and recoveries and the vaccination program. This massive information flow requires students to be able to distinguish between fact and opinion, reliable and unreliable information as well as to detect biased information and fake news (Delport, 2023; Suarez-Alvarez, 2021). Students are, therefore, expected to develop their own critical interpretations and evaluations of information context (Yilmaz et al., 2023). Additionally, they should enhance their skills in data visualisation and constructing data-based arguments (Bailey & McCulloch, 2023; Engledowl & Weiland, 2021; Lee et al., 2022) rather than solely relying on external sources. In other words, statistical literacy (SL) is becoming increasingly crucial for all students to become informed and well-educated citizens (Budgett & Renelle, 2023; Budgett & Rose, 2017; Büscher, 2022a; Delport, 2023; Franklin, 2021; Johannssen et al., 2021).

Nonetheless, most adults tend to misinterpret statistical information (Dahlstrom-Hakki & Wallace, 2022; von Roten & de Roten 2013). Additionally, statistical knowledge is lacking among many high school students in the majority of developing countries. This issue is highlighted by the Programme for International Student Assessment (PISA) which evaluate the knowledge of 15- to 16-year-old students. PISA results indicate that students from developing countries have consistently performed poorly on the uncertainty and data subscale for the past two decades. In the PISA 2003 report (OECD, 2004), the lowest-ranking countries for this subscale were exclusively developing countries where 50% to 80% of their students scored at and below Level 1, the lowest proficiency level. Level 1 of PISA on this subscale suggests that students are only able to locate specific data values from a simple representation, while below Level 1 is an additional level to accommodate students who could not achieve level one. Two decades later, in the PISA 2012 and 2022 tests, students from these countries exhibited minimal improvement on the uncertainty and data subscale (OECD, 2014, 2023). Moreover, students from certain developing countries that participated in the PISA test for the first time also displayed notably low performance (OECD, 2014, 2023). This consistent underperformance of 15- to 16-year-old students from developing countries over the past two decades raises concerns about their proficiency in solving data-related problems and whether students at this age are sufficiently statistically literate.

Given that all students are expected to become statistically literate citizens after leaving school (Büscher, 2022a; Franklin, 2021; Franklin & Bargagliotti, 2020; Gal, 2002; Watson & Callingham, 2020), students in their final year of formal schooling should demonstrate sufficient SL. However, high school students' SL seems to have changed little in the ten years between the various studies, and their comprehension is predominantly non-critical (Callingham & Watson, 2017). Moreover, there is a scarcity of data concerning students' SL during their final year of schooling, particularly in developing countries. Studies on statistics education involving upper high school or final-year students have primarily been conducted in Western contexts (e.g., Budgett & Rose, 2017; Dierdorp et al., 2017; Gil & Gibbs, 2017) with insufficient studies conducted in non-Western contexts (e.g., Aoyama, 2007; Hafiyusholeh et al., 2018; Sharma, 2014). Considering these findings, this study intended to examine students' SL in the final year of schooling from non-Western and non-developed countries. There were three purposes for involving students from the final year of schooling: to contribute to the limited theory in the field, to reveal the SL of these future citizens of developing countries and, most importantly, to investigate the progress these students have made compared to those from lower grades.

In addition, the distribution of SL studies indicates that researchers are strongly encouraged to conduct SL studies of students in Asia, particularly Indonesia. Marchy and Juandi (2023) conducted a systematic literature review to determine the trend of SL studies worldwide, whereas Carel and Juandi (2023) conducted a systematic literature review to determine the trend of SL studies in Indonesia. Marchy and Juandi's (2023) analysis of 70 English-language publications (from 1980 to February 2023) revealed the distribution of SL studies by publication year, level of education of participants and country of study. The results showed that, for 44 years, most of the articles were not published until 2003, with the greatest number of publications occurring in 2017 (Marchy & Juandi, 2023). Participants in these studies ranged in level of education from elementary school to university, with the majority of studies being conducted at the secondary level, followed by the university level. Regarding country of publication, studies on SL have been conducted predominantly within the United States, with a total of 21 publications. Subsequently, Australia has contributed significantly to this field with 16 publications. In contrast, the scholarly output in Indonesia has been notably limited, with a mere two publications identified. Carel and Juandi (2023) then identified and analysed 40 SL studies in Indonesia that were written in Bahasa (Indonesian language) or English and published from 2017 to 2022. The results showed that most articles were published in 2021, conducted at university levels and junior high school, and carried out on the island of Java, specifically in the provinces of West Java and East Java. However, these publications focused on instruction rather than assessment.

Based on a review of the statistics education literature, numerous assessment frameworks have been developed and various studies have been conducted to assess high school students' SL. However, most studies have investigated the SL of students within one grade or the SL of students majorly from the same grade (e.g., Çatman Aksoy & Işıksal Bostan, 2021; Koparan & Güven, 2015; Mullis et al., 2012; OECD, 2004, 2014, 2023; Pfannkuch, 2005). Only a limited number of studies have investigated the development of students' SL across grades. First, Callingham and Watson (2017) conducted a longitudinal study in Australia with students in Years 5 to 10. They discovered that while there was no growth in SL from Years 5 to 6 and from Years 9 to 10, there was growth throughout the transition from primary to secondary school (Year 6 to Year 7). Second, Aoyama and Stephens (2003) conducted a study in Japan comparing Years 5 and 8 students. They reported an improvement in students' SL. However, it is unclear whether that improvement can be attributed to formal statistical education due to statistics has not been covered very extensively between the two grades. Instead, they claimed that this improvement might be linked to general cognitive development, including students' exposure to data-based information both inside and outside the classroom. Third, Yolcu (2014) investigated the SL of Turkish students across Years 6 to 8, considering factors such as grade and gender. She discovered no significant grade-related difference in SL across Years 6 to 8. Further, she claimed that this lack of difference might be due to the adjacent grades and the spiral curriculum in middle school mathematics—wherein the same topic is repeated in multiple grades and occasionally with the slightly increasing levels of difficulty and competence (Snider, 2004; Wang & McDougall, 2019). Even though each of these studies investigated the development of SL by grades, none were conducted in developing countries with students in their final year of schooling.

To fill this gap in the literature, the present cross-sectional study investigated SL in Indonesian Year 9 and Year 12 students. Indonesia was selected because it was one of the developing countries performing poorly on the uncertainty and data subscale in PISA 2003, 2012 and 2022 tests (OECD, 2004, 2014, 2023) and underrepresented country in SL studies (Carel & Juandi, 2023). In the PISA 2003 test, Indonesia was ranked 38 out of 40 participating countries, with approximately 72% of students scoring at and below Level 1 on the uncertainty subscale; in the PISA 2012 test, Indonesia ranked 63 out of 64 participating countries, with approximately 73% of students scoring at and below Level 1; and in PISA 2022 test, Indonesia ranked 70 out of 77 participating countries. For the present study, Year 9 and Year 12 students were chosen because these cohorts represent Indonesian students participating in the PISA test and the final year of schooling, respectively.

Although PISA provides important results (e.g., the levels of Indonesian students on the uncertainty and data subscale, the trend over the years of participation and the types of items used to assess students' SL), the report does not address current assessment needs (e.g., students' responses when encountering data-based claims, the effects of grade levels on students' SL and students' challenges when responding to information containing statistics). The current study utilises a combined quantitative and qualitative cross-sectional design to assess the SL of Year 9 and Year 12 students. This assessment employs a novel SL framework tailored to equip students with the skills required to respond to statistical information. The evidence for the reliability and validity of the framework and instrument was gathered throughout the development process and piloting. A test and an interview were then administered to Indonesian high school students from different cohorts. It was then possible to compare the SL of students from different cohorts and investigate the challenges they faced as well as their understandings, in addition to revealing their responses and SL levels.

1.2 Personal Motives for Choosing Assessment on Statistical Literacy (SL)

In the beginning of this study, the researcher identified two personal motivations for conducting a study on SL assessment, and one more motivation emerged midway through. These motivations took into account both the researcher's previous research and his intention to expand the scope of his study interests as a mathematics researcher. More significantly, this work represents a form of dedication to develop mathematics (and statistics) education in Indonesia.

The researcher's initial motivation stems from his past research experience, which was mostly focused on developing an instructional theory of school mathematics. He has focused his prior research on designing lessons for students to interpret data from bar and line graphs. It would be thorough if he began concentrating on an assessment study to complement his prior research experiences. Moreover, an international test such as PISA revealed that Indonesian students' performance was exceptionally low, even after the new curriculum 2013 (K13) was introduced in 2013. Therefore, he felt personally responsible to find the most effective way to assess students' understanding, particularly on their SL.

The low SL shown by Indonesian high school students is his second motivation. He was worried about how Indonesians, in the future, might react to information that is based on data. Whereas he mostly studied elementary school students in his earlier research, this time he concentrated on high school students. The SL of these high school students both reflects and predicts the level of responses of Indonesians to statistical information. Thus, he wanted to investigate the current state of SL of Indonesian high school students across grade levels and what can be done about it.

Along with the first two motives, the COVID-19 cases served as an additional motive for him to carry out this study. The COVID-19 pandemic encouraged everyone to possess sufficient SL, including the researcher. The COVID-19 pandemic taught him that being literate in one area of statistics does not imply proficiency in all areas of statistics. As soon as he saw the massive amounts of data related to the COVID-19 pandemic, he frequently experienced confusion and illiteracy. Furthermore, it was difficult to make personal decisions in the middle of a chaotic situation in Indonesia. Hence, he wanted to become a more proficient researcher in the field by researching this topic, in addition to advancing research in statistics education.

1.3 Research Aims and Questions

Considering that Indonesian students performed poorly in solving problems containing data, this study aimed to investigate the SL levels of Indonesian high school students across grade levels. Given the lack of comprehensive information on SL from PISA and the Trends in International Mathematics and Science Study (TIMSS) reports, Gal (2002) recommended an in-depth qualitative investigation into students' thought processes and comprehension. This recommendation remains relevant due to the limited number of studies focusing on this area over the years. Consequently, the present study also aimed to investigate students' challenges and understandings when responding to data information. This investigation focused particularly on the four skills—interpreting, communicating, evaluating and decision making— and three components associated with SL—text and context, representation and statistical-mathematical knowledge— (as detailed in Chapter 2). An instrument and an interview protocol were designed to achieve the aims (refer to Chapter 4). The students' written responses from the test were intended to reveal their SL levels, challenges and understandings, while the students' spoken responses from the interview were expected to provide additional information on their challenges and understandings.

To ensure comprehensive insights, participants were drawn from diverse backgrounds, in addition to different grade levels (as detailed in Chapter 5). This approach not only enables a comparative analysis of SL levels but also contributes to addressing educational disparities in Indonesia. The following questions then guided this study:

1. What levels of SL do Indonesian high school students possess?
2. Are there any significant differences in Indonesian high school students' SL based on their demographic backgrounds (i.e., grade level, gender, school type, school status or city of origin)?
3. How do the challenges students encounter in comprehending the three components of SL affect their abilities to respond to statistical information?
4. How do students' understandings of the three components of SL influence their abilities to respond to statistical information?

1.4 Research Significance

This study has three research significances. First, it expands the scope of studies in this field to a developing country that is contextually and culturally different to Western countries. This study is significant because it focuses on a population often underrepresented in research, Indonesian high school students. This study is also significant because it provides a better understanding of how Indonesian high school students process mathematics tasks requiring statistical thinking, which is useful given their low performance on international assessments. In particular, the cross-sectional design employed in this study facilitates cohort comparison, enabling the assessment of SL progress between Year 12 and Year 9 students.

Second, this study aimed to investigate the SL levels of Indonesian high school students by using an innovative assessment framework consisting of four SL skills to describe the interaction of three SL components. From a theoretical viewpoint, the interaction of three SL components determines students' SL levels. Additionally, this study deepens our understanding of this interaction when students encounter challenges and exhibit understandings while responding to data-based information.

Third, this study is valuable for Indonesian education stakeholders. The investigation into the various SL skills provides detailed insight into Indonesian students' skill-based responses. Curriculum developers can benefit from the findings of this study by enhancing the existing standard competencies students need to achieve. Then, the statistical content in mathematics textbooks can be adjusted in accordance with the changes in standard competencies. Textbook designers become informed about the typical tasks and expected levels of responses required for standard competencies. Further, mathematics teachers will gain insights to effectively design SL assessment instruments. Researchers in mathematics education interested in studying students' SL can replicate or extend this study to other regions of Indonesia or different contexts. A review of 40 SL studies in Indonesia by Carel and Juandi (2023) revealed a lack of similar studies, underscoring the crucial significance of replicating and extending this research.

1.5 Overview of Thesis Chapters

This section provides a brief overview of how this thesis is structured. There are eight chapters: 1) introduction, 2) literature review, 3) study context, 4) instrument development, piloting and refinement, 5) methodology, 6) students' SL, 7) students' challenges and understandings when responding to statistical information and 8) study summary.

Chapter 1 introduces the rationale for assessing Indonesian high school students' SL and presents its significance. In Chapter 2, a comprehensive review of literature on SL definition and assessment in the high school context is conducted, demonstrating the necessity for a new SL assessment framework. Using the newly proposed SL assessment framework as a lens, Chapter 3 provides the study context by reviewing the Indonesian curriculum, mathematics textbooks, teachers' demography and pedagogy, and both *Ujian*

Nasional (UN; Indonesian National Examination) and international mathematics tests to reveal the opportunities that existed to conduct an assessment study with Indonesian high school students.

Chapters 4 and 5 describe the methodology. In Chapter 4, emphasis is placed on instrument development and piloting, while Chapter 5 details the research design and overall methodology. In Chapter 4, three instrumentations are discussed: the test items, scoring guide descriptors and interview protocol. This chapter establishes the applicability of the SL assessment framework. Chapter 5 describes the theoretical foundation of the cross-sectional study, along with participant details, data collection and analyses.

Chapter 6 presents the results of the quantitative and statistical analyses, which were designed to address the first and second research questions, those about students' SL levels and SL differences. Chapter 6 also presents a discussion of the quantitative findings, considering 1) the differences in students' SL, 2) the future participation of Indonesian students in society and 3) the applicability of the SL assessment framework for mathematics teachers. Chapter 7 presents the results of the qualitative analyses, which were designed to address the third and fourth research questions, those about students' challenges and understandings. This chapter also presents a discussion of the challenges faced by students and their understandings of the three SL components, addressing 1) the interrelationships of the three components of SL and 2) a call for an SL environment. Finally, Chapter 8 provides a summary of the four research questions and their corresponding findings. It further explores the limitations and offers implications for future studies and statistics teaching.

Chapter 2: Literature Review

This chapter discusses prior research concerning the statistical literacy (SL) of high school students. Section 2.1 reviews the SL definitions in the existing literature to identify the essential skills needed by students to be statistically literate. Further, Section 2.2 identifies certain knowledge that can contribute to students' SL. Section 2.3 reviews the existing assessment frameworks to comprehensively understand how high school students' SL has been assessed and what kind of levels that characterise students' SL. Section 2.4 reviews studies that focused on between-group disparities in students' SL. Section 2.5 presents factors that influence students to encounter challenges and demonstrate understandings in SL. Section 2.6 summarises and explains the need for a new SL assessment framework to assess high school students as consumers of information containing statistics. Finally, Section 2.7 provides a chapter summary.

2.1 Defining SL

Currently, and since the 2000s, there has been little consensus on the definition of SL (Bailey & McCulloch, 2023; Budgett & Pfannkuch, 2010; Budgett & Renelle, 2023; Budgett & Rose, 2017; Büscher, 2022a; Gal, 2019; Helenius et al., 2020; Jureckova & Csachova, 2020; Sabbag et al., 2018; Schield, 2017; Sharma, 2017). Nevertheless, existing definitions can be classified into two major perspectives: those of data producers and those of data consumers (Budgett & Rose, 2017; Gal, 2002; Wild, 2017). In addition, the definition can involve both perspectives in a given context.

The data producer perspective is centred on an individual's proficiency in employing statistical processes, such as formulating question and gathering data (Franklin, 2021; Franklin & Bargagliotti, 2020; Franklin et al., 2005; Perez et al., 2021) as well as constructing graphs and tables (Koparan & Güven, 2015; Pallauta et al., 2021). Conversely, the data consumer perspective examines a person’s ability or skill in understanding and assessing statistical information (Cui et al., 2023; Wallman, 1993), making personal daily choices based on data and information (Budgett & Renelle, 2023; Franklin et al., 2005), critically evaluating data-based information and relevant media-presented news (Budgett & Rose, 2017; Büscher, 2022a; Guler et al., 2016; Koga, 2022b) and interpreting, critically evaluating and communicating statistical results from diverse sources (Cui et al., 2023; Gal, 2002; Sutherland et al., 2022). Altogether, four essential skills have been identified for data consumers: **interpreting, evaluating, communicating and decision-making**.

In the 21st century, high school students are required to demonstrate these essential data consumption skills, as part of SL, since they are becoming increasingly active consumers of statistics. SL has become a crucial component of their daily lives, as indicated by various studies (Budgett & Renelle, 2023; Budgett & Rose, 2017; Dahlstrom-Hakki & Wallace, 2022; Ludewig et al., 2020; Sutherland et al., 2022). The core of SL is a critical response to statistical information (Callingham & Watson, 2017; Büscher, 2022a; Sharma, 2017; Sharma, 2018b), which is regarded a higher-order competence (Koga, 2022a). Students who possess critical thinking are better equipped to seek for credible information sources (Hoffrén, 2021), guiding them to be statistically literate citizens, statistical citizenship, for the world to come (Budgett & Rose, 2017; Wild, 2017). As they transition into adulthood, data consumption skills empower them to actively engage in communities to make sense of governmental reports and policy decisions, understand political polls and societal data, understand scientific trends and phenomena and perform job-related duties (Sutherland et al., 2022; Weiland, 2019).

However, previous studies have primarily examined the skills of data consumers in isolation, rather than considering them holistically. For example, a study by Aoyama and Stephens (2003) focused on students’ graphical interpretation levels, while a study by Sharma

(2006) focused on students' understanding of graphs and tables. Further, Guler et al. (2016) focused on assessing Year 8 students' critical evaluation of data presented in newspapers, using 10 critical questions developed by Gal (2002). Additionally, Budgett and Rose (2017) focused on evaluating skills by designing a teaching approach to facilitate students' ability to critically evaluate media reports. In contrast to addressing the four fundamental skills individually, the current study integrated them. In other words, this study demonstrates that statistically literate students are those having the skills of interpreting, communicating and evaluating statistical information and using it in making informed decisions. This necessitates the clarification of the nature of these respective skills, which are explained further in subsequent sections of this chapter.

2.1.1 Interpreting

In the studies of students' SL, interpreting is considered the most frequently used skill. However, the definition of this skill varies across different studies. The GAISE (Guidelines for Assessment and Instruction in Statistics Education) and LOCUS (Levels of Conceptual Understanding in Statistics) frameworks define interpreting as the ability to look beyond collected data to solve statistical problems (Franklin, 2021; Franklin & Bargagliotti, 2020; Franklin et al., 2005; Perez et al., 2021; Whitaker et al., 2015). Interpreting also relates to students' abilities to identify trends and make predictions from graphs (Mooney, 2002). This aligns with Curcio's (1987) terms 'reading of the data', 'reading between the data' and 'reading beyond the data.' Patahuddin and Lowrie (2019) applied this Curcio's level to assess teachers' interpretation of a context-based line graph. Based on the PISA 2022 framework (OECD, 2018), interpreting entails understanding information from graphs or tables and applying mathematical and statistical results in real-world contexts. Considering these distinct definitions, interpreting can be classified into basic and advanced data sense-making.

At its core, interpreting involves understanding and utilising basic statistical languages and symbols within specific contexts (Ben-Zvi & Garfield, 2004; Franklin et al., 2005; Garfield et al., 2010). For example, students need to demonstrate an understanding of the mean, mode and median along with their corresponding symbols and meanings within a specific context. An indicator of understanding is that students can use those statistical terms to respond to problems (Chance, 2002) or statistical results (Wallman, 1993). When it comes to basic graph or table reading, Aoyama and Stephens (2003) use the term 'interpreting' to refer to this basic competency, while Sharma (2006) includes Curcio's 'reading of the data' level (i.e., the reading of information that has been presented clearly).

Once they have more advanced interpreting skills, students are expected to develop more critical responses to graph or table reading. According to Rumsey (2002), interpreting statistical data involves an in-depth comprehension of the meaning of the data. Competence in this area is demonstrated by the ability to extract qualitative meaning from data that are frequently presented quantitatively (Aoyama & Stephens, 2003). It also refers to students' ability to read information that is implicit—not explicitly displayed—in a table or graph (Sharma, 2006). Sharma (2013a) classifies this ability as 'reading beyond the data'. This skill involves identifying or extracting the qualitative meaning of statistical concepts, such as average or trend, from the graphed data.

In summary, interpreting stands as a pivotal skill in SL, bridging data and insights. Its dynamic definitions encompass fundamental comprehension and advanced analysis, enabling students to navigate complex information and derive valuable meaning. This skill also equips students to effectively communicate their understanding to others.

2.1.2 Communicating

In a data-driven world, statistical survey results and data representations often become publicly available to citizens. The amount of statistical data that is available through online media is growing, and the internet is also continually developing new kinds of virtual communication displacing actual communication (Hoffrén, 2021; Marchy & Juandi, 2023; Wild, 2017). Citizens often express their reactions to statistical information or data representation through formal or informal communication and through spoken or written words. This information sharing (communicating) involves sharing or discussing reactions, such as an understanding or opinion of the meaning of data-based information with others (Gal, 2002; Sharma et al., 2011). As members of the data consumer community, students play a vital role in ensuring accurate information dissemination and preventing the spread of misinterpreted data. More importantly, such communication needs to be conducted effectively so that others, including statisticians and non-statisticians, can understand the information (Gal, 2002; Krishnan, 2015). Otherwise, there would be widespread (mis)representations and (mis)interpretations (Engledowl & Weiland, 2021).

Due to the importance of communicating quantitative information, a specific type of question needs to be designed to put students in a position to share or discuss their understanding. Notably, exemplary instances of such questions can be found in the International English Language Testing System (IELTS) Writing Task 1, which assesses the test taker's ability to describe and explain information using data. The test takers are commonly provided with a context and with data in a graph, chart, table or a combination of two, followed by questions, such as 'write a report that describes the information' or 'summarise the information by selecting and reporting the main features' (Freimuth, 2016). The test takers need to select the best main features by describing and comparing the data using their own words. They need to analyse and synthesise the most relevant data to include in the writing, provide an adequate written response to the visually represented data and choose appropriate words (Valentina, 2016).

Writing a well-structured response might be challenging for high school students, but observing students' written responses could inform observers about their critical thinking. The main aim of testing high school students' skills in communicating quantitative information is to find out how critical they are in communicating the most relevant data. Written responses should be written for others to understand easily and should include numeric information, highlight trends, make comparisons, show relationships and include the most significant data (Sherington et al., 2004). Therefore, the students need both linguistic competence and graphical competence (Rotaru, 2018). Further, Rotaru (2018) stated that although students—in Romania—from all grades are familiar with various forms of graphical representation, they do not receive enough training in communicating information presented in graphs. This situation might be comparable to that of Indonesian students, who have been exposed to various types of data representation since elementary school, but statistical instruction mainly focusses on number and formulas, instead of critical graph interpretation (Padmi, et al., 2018). Consequently, evaluating Indonesian high school students on this skill is essential for determining their ability to communicate data-related information.

In summary, communicating reactions to statistical information stands as the second crucial skill in SL. Mastering this skill helps to stop the spread of inaccurate information throughout the community. However, students need to be more cautious while sharing their viewpoints when there are claims in the information that has to be communicated. They should be prepared to critically evaluate such claims.

2.1.3 Evaluating

Data-based information has the potential to be misleading. Coping with such misleading information requires students to possess a capacity to evaluate statistics-based claims or arguments. Statistics-based claims are becoming more pervasive than in the past (Weiland, 2019) and even high-quality data can mislead the readers (Wild, 2017). These misleading claims, which have appeared in online and printed media as well as emerged in informal conversations among citizens, are frequently accompanied by evidence. However, statistical claims that appear in the media may not be free of bias and the analysts and communicators may intentionally misreport and misrepresent it (Delport, 2023; Gal, 2002; Wild, 2017). Additionally, claims emerging from an informal conversation might lack empirical support. Therefore, students must demonstrate their capability to challenge claims or arguments using data (Brown et al., 2010; Franklin, 2021; Franklin & Bargagliotti, 2020; Sharma et al., 2011). They need to position themselves as critical data consumers to actively participate in society (Marchy & Juandi, 2023) and to prevent themselves from being misinformed.

When challenging a statistical claim, students' arguments should include a reasonable and critical question using the data upon which the claims were based. Gal (2002) provided a list of critical questions readers can use to evaluate or argue against statistical claims. Guler et al. (2016) further classified these critical questions into categories, such as consistency and data presentation. A typical question in the consistency category is, 'Are the claims here sensible and supported by the data?' By proposing such a question in their minds, students can check the consistencies or inconsistencies of a claim against the provided statistics, tables and graphs and even critically assess the process through which the claim was developed. Consistency also refers to checking the reliability of the evidence and how this statistical evidence relates to the claim (Koga, 2022a; 2022b). It is necessary to examine whether the evidence is sufficient to support the claim (Koga, 2022b). Further, a typical question in the data presentation category is, 'Is a given graph drawn correctly?' Students asking this critical question tend to focus on whether the graph was drawn correctly and why it was drawn in a particular way. Students posing this graph-related question demonstrate ability to make judgements about validity of claims based on graphs (Sharma, 2013a).

In addition to the two categories outlined above, critical questions can also be posed in relation to the data set's context, statistical content and the limitation of the given data set (Delport, 2023; Koga, 2022b; Yilmaz et al., 2023). Statistics is number in context (Cobb & Moore, 1997; Shaughnessy, 2007), so it is necessary to determine whether the existing claim is relevant to the context from which it arises. Statistical concepts used as evidence in the claim must also be checked from different perspective. Students need to be able to use statistical knowledge in different ways and in different situations (Jureckova & Csachova, 2020), in order to contest a claim. However, to properly evaluate statistical claim, students should provide only the necessary evidence and not more than is required (Woodard et al., 2020). Additionally, they should provide a counterargument supported by good mathematical writing, allowing the reader to readily follow the logic of each step (Woodard et al., 2020). In the case of the data are insufficient to support the existing claim, there is no reason to place excessive trust in the information (Delport, 2023).

It is then important to use a test to assess students' ability to evaluate statistical information. Such a test can reveal the quality of their critical arguments and their various approaches to challenging a claim. Three possibilities may arise from the students' responses (see, e.g., Gal, 2002; Guler et al., 2016; Rumsey, 2002; Wallman, 1993; Weiland, 2017). First, the student may construct a logical argument and collect evidence to challenge the claim. Second, the student may construct an argument and collect evidence to support the claim. Third, the student may find no evidence to challenge the claim, so they accept it without question.

In summary, critically evaluating statistical claims or arguments stands as the third crucial skill in SL. This skill encompasses students' ability to challenge claims made by others, utilising relevant evidence from the data sources of those claims. It is then anticipated that students should also be able to include relevant evidence to support their arguments when they must draw their own claim (decision) from the data provided.

2.1.4 Decision-Making

Data-based information is ubiquitous and is frequently designed to persuade data consumers; therefore, decision-making using statistics is the responsibility of all individuals. Statisticians and official institutions might handle decisions involving big data and advanced analytics for societal policies. However, at an individual level, individuals also need to make informed decisions using quantitative information and statistical arguments (Callingham & Watson, 2017; Cui et al., 2023; Franklin & Bargagliotti, 2020; Weiland, 2017). All individuals, including students, need decision-making skills for daily choices (Budgett & Renelle, 2023) and personal decisions (Berndt et al., 2021; Koga 2022a; Krishnan, 2015). As data consumers, high school students can develop these skills by making decisions based on statistics concepts they learn, such as measures of central tendency and spread as well as various types of representation. Using real-world problem-solving involving these ideas, for example, students can decide which measurement centre is best or more suitable.

In addition, students are expected to demonstrate statistical reasoning and problem-solving skills when confronted with real-world decision problems. What to buy, which universities to apply for and who to vote for are examples of real-world decision problems (Cui et al., 2023). Critical statistical thinking is then required as part of this logical process (Guler et al., 2016). For example, students must be able to select appropriate statistics for determining which product to choose if similar products have different prices or qualities. Existing research found that a price reduction (Cecere et al., 2018) or product warranty (Li et al., 2020) were often important factors that consumers needed to consider before buying a product. Additionally, social media extensively affects personal and managerial decision-making (Power & Phillips-Wren, 2011). For example, travellers use social media to gather information and reviews related to destinations, transport and accommodation from other travellers before deciding on a place to visit (Dwityas & Briandana, 2017). Reviews from travellers are also used by managers to promote their destinations to more travellers. Citizens (including students) need to be aware of this cycle and to consider all the options to make informed decisions.

Due to the importance of decision-making skills, high school students need to be assessed on this skill. Students' ability to support their arguments with data-based evidence is one of the main goals of teaching statistics (Woodard et al., 2020). However, tests of this skill are rarely found in the existing assessment studies of SL. A good example of a question that assesses decision-making skills can be found in Sharma et al.'s (2012) 100-Metre Race item. Sharma et al. (2012) did not label this 100-Metre Race item as a decision-making test. However, in this item, students were asked to select one of three runners based on the times recorded for seven races, and this tested their skills to make decisions based on statistical information.

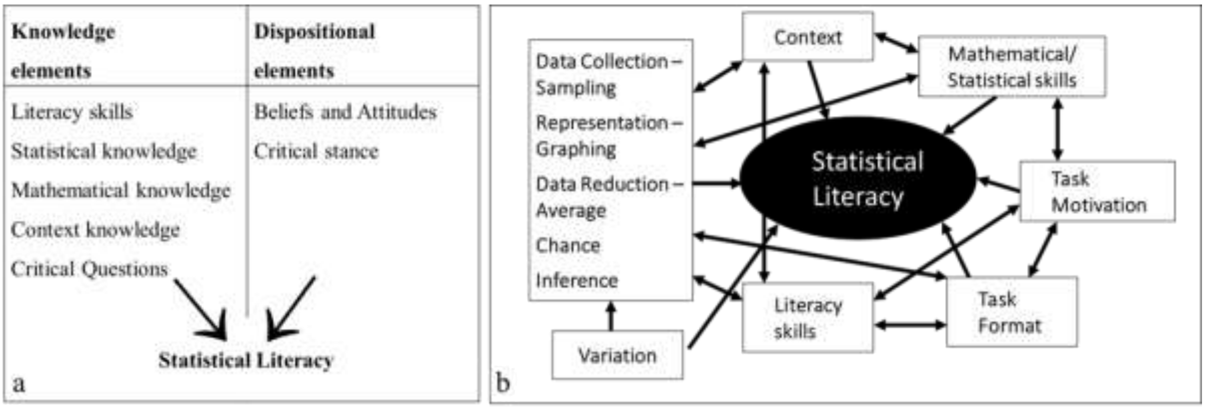
In conclusion, the fourth crucial skill in SL is decision-making. Students must be able to support their decisions with pertinent evidence in this data-driven world. To make the right decision, they need to apply their knowledge of statistical concepts as well as various types of representation they learn at school. In addition to decision-making skill, the other three SL skills—interpreting, communicating and evaluating—also require students to apply statistical knowledge they learn at school. This type of knowledge is covered in the following section.

2.2 Components Contributing to Students' SL

Responding to statistical problems is a complex process involving the activation of various forms of knowledge. Students' ability to execute the four SL skills to respond to statistical information is influenced by their knowledge of the contributing components. Gal (2002) highlights two fundamental factors that contribute to the SL of adults: knowledge elements and dispositional elements (depicted in Figure 2.1a). The knowledge component consists of five subcomponents: literacy skills, statistical knowledge, mathematical knowledge, contextual knowledge and critical questions. The dispositional component incorporates two subcomponents: beliefs and attitudes, and critical stance. Gal states that these combined components contribute to the development of SL in adults. Watson (2006) likewise contends that different type of knowledge greatly influences students' SL. Examples of these components include literacy, knowledge of context, mathematics, statistics and task format. Watson further agrees with Gal that these contributing components do not support students' SL as separate entities; instead, the components are interconnected.

Figure 2.1

Components of Statistical Literacy (SL)



Note. (a) Components of SL in adults (Gal, 2002, p. 4); (b) Components of SL in students (Watson, 2006, p. 248).

The interrelationship among these contributing components in relation to students' SL is clarified by Watson (2006), as illustrated in Figure 2.1b. The understandings of these components by students are interdependent. For example, students' appreciation of context will determine and be determined by their literacy and their statistical (as well as mathematical) knowledge (depicted by the directional arrows in Figure 2.1b). Further, those three components will influence and be influenced by students' competence in graphing; for which graphical competence is vital for all individuals, including students (Olande, 2014). Because statistical items in international tests (such as PISA and TIMSS) and other SL-assessing studies frequently consist of features relating to those components, students' levels of ability in each component need to be characterised.

Even though there are many assessment studies on students' understanding of data, limited studies classify students' responses relating to each of the contributing components into a hierarchy. Yotongyos et al. (2015) classified students' levels of knowledge of contributing components into the categories of high, moderate and low based on the students' mean score of the 7-point Likert scale. These levels determined a student's overall level of knowledge components. Although this is an innovative viewpoint, the characteristics of each level were not clearly defined. In comparison, Koparan and Güven (2015) developed descriptors for data representation across Watson and Callingham's (2003) six increasing levels: idiosyncratic, informal, inconsistent, consistent non-critical, critical and critical mathematical. However, the descriptors are limited to graph reading and construction, whereas Watson and Callingham's (2003) six levels essentially provide thorough descriptors that involve representation, statistical-mathematical knowledge and contextual knowledge. Moreover, existing SL-assessing studies also primarily focus on assessing students' levels in general rather than their specific levels of the contributing components. Considering these findings, the present study assessed students' levels for each component as the foundation of determining students' SL levels.

All of the knowledge components outlined by Gal (2002) and Watson (2006) are essential for assessing the SL levels of high school students. Inadequate understanding of any of these components could leave students inadequately informed. However, there is a subset of similar components that Gal (2002) and Watson (2006) identified. Due to their similarities, these SL components were adjusted slightly. According to Gal (2002) literacy skills demand text and graph comprehension, which this study distinguished. Text comprehension was grouped together with the contextual understanding—as part of the reading skills (Yilmaz et al., 2023)—and renamed *text and context*. Graph comprehension that examines different kinds of processes to make sense of graphs and tables was renamed *representation*. Statistical knowledge and mathematical knowledge were grouped as *statistical-mathematical knowledge*. Further details of each component are explained in the following sections.

2.2.1 Text and Context

Data consumers should understand the plain text information that provides context for statistical data in digital and printed media (Gal, 2002; Koga, 2022b). Data is a number with a context (Cobb & Moore, 1997; Shaughnessy, 2007), and when used properly in context it can be invaluable (Franklin & Bargagliotti, 2020). The consideration of context is crucial in the discipline of statistics (Bailey & McCulloch, 2023; Sharma, 2013b; Weiland, 2019) to understand the narrative reflected in the data (OECD, 2018; Sharma et al., 2011; Yilmaz et al., 2023). Context covers various life and global aspects—social, economic, environmental, personal, educational, occupational, public and scientific (Franklin, 2021; Gal, 2019; OECD, 2018). Additionally, context provides students with strategies to solve mathematical problems (Van den Heuvel-Panhuizen, 1996) and motivates procedures to solve statistical problems (Gal, 2002, 2019). Considering the importance of context in tackling statistical problems, students, as data consumers, should possess sufficient levels of contextual knowledge.

In the existing SL assessment studies, all statistical problems are designed to have a specific context. According to Gal (2019), the context involved should be authentic or naturally occurring in the real world (not invented or fictitious). However, if students collect their own data, the data can be real even if the context is invented. The use of real or fictitious contexts is a point of debate among experts. Wijaya et al. (2014) agrees that the contexts should not necessarily be limited to real-world settings; the important consideration is that situations can be created that relate to students' common-sense understanding. For instance, students participating in PISA and TIMSS might be disoriented when encountering unfamiliar contexts; however, they are still expected to solve problems contextually using their common sense although the context exceeds their familiarity.

Further, understanding data-based information requires students to be able to process text (Gal, 2002). Students might be unable to examine the context and societal issues that underpin the information if they are not familiar with the surrounding text or background information (Delport, 2023). The textual information might consist of statistical or other terminology specific to a certain context. Students' responses to this textual-contextual information vary from personal to critical engagement (Sharma, 2013a), as outlined in SL level descriptors within frameworks like Watson (2006), Watson and Callingham (2003), PISA and TIMSS. Text comprehension enables students to make sense of unfamiliar context from outside school (Sharma, 2018b). Furthermore, knowledge of data's context—often presented in text—helps in explaining the data, gaining insights, identifying pertinent information and justifying a claim (Yilmaz et al., 2023). Consequently, students' engagement with the text and context component when solving data-based problems is investigated in this study.

2.2.2 Representation

Graphical competence is vital for all individuals, including students, and considered part of SL (Kemp & Kissane, 2010; Sharma, 2006). This importance is driven by the prevalence of data representations presented in the media (Arteaga et al., 2021; Bailey & McCulloch, 2023; Olande, 2014) and in numerous areas of society for general public consumption (Bursal & Yetiş, 2020; Jureckova & Csachova, 2020; Patahuddin & Lowrie, 2019). The advances of technology have increased the possibility of students encountering, at least, two types of representation (graphs and tables) in various media and contexts (Sharma, 2013a). Thus, understanding how to make sense of data presented in graphs and tables is equally important. This competence is among the most important skills in today's world since representations can effectively visualize the rapidly expanding amounts of data (Bursal & Yetiş, 2020). Considering its importance, there are, at least, four other terms have been used to denote graphical competence from previous studies: graph literacy, graphicacy, graphing ability and graph sense (Ludewig et al., 2020). Nevertheless, henceforth, the term graphical competence is used to refer to students' ability in making sense of data displayed in both graphs and tables.

Due to the importance of graphical competence to support students' SL, existing studies of students' SL levels describe the graphical competence required at each of their levels. Depending on the specific information required from a statistical graph, various levels of challenging questions can be formulated (Arteaga et al., 2021). Similarly, students' responses to graph and table-based problems requiring critical thinking can vary in their level of complexity. In the four levels of the TIMSS, students' graphical competence is described as progressing from the ability to understand basic representations (tables and graphs), the ability to read and interpret data representations and then to the ability to reason in a variety of problem situations and make generalisations (Mullis et al., 2009). Likewise, in PISA's six levels, students' graphical competence ranges from reading the information in a simple graph and table to interpreting, evaluating and critically reflecting on complex statistical data and communicating reasoning (OECD, 2013a). Watson and Callingham (2003) and Callingham and Watson (2017) also described the level of ability to interpret data representations that is required at each of their six hierarchical levels, from the idiosyncratic to the critical mathematical levels. Further, Koparan and Güven (2015) developed descriptors for data representation at each of Watson and Callingham's six increasing levels to measure students' data interpretation capabilities.

Some previous studies also assessed students' graphical competence and provided steps for the interpretation of tables and graphs. Those studies focused on students' graphical and tabular interpretation skills (Aoyama & Stephens, 2003; Sharma, 2006) or on students' critical evaluation of data presented in the newspaper such as in Guler et al. (2016) who used 10 critical questions developed by Gal (2002). When interpreting tables and graphs, students were expected to fluently extract and use information from those representations (Ludewig et al., 2020) as well as include numerical interpretations and opinion statements (Moritz, 2003). In line with Moritz (2003), Kemp and Kissane (2010) developed a five-step framework for helping students to interpret data presented in tables and graphs, namely:

1. Students need to examine all the features of the graph or table (i.e., title, axes, headings, legends, footnotes, source) to discover the context.
2. Students have to find what the numbers represent (e.g., by looking for the largest and smallest values in one or more categories to obtain an impression of the data).
3. Students need to find the differences in the values (e.g., the differences between the data in rows or columns, the changes of data over time, and the comparative values of data within a category).
4. Students must continue to identify where differences occur (e.g., using information from Step 3 to make comparisons between two or more categories or timeframes).
5. Students must assess why those differences occurred by looking for reasons for the relationships in the data and relating them to the context.

This five-step framework successfully helped primary- to tertiary-level students interpret tables and graphs. Particularly, the fifth step is consistent with Arteaga et al. (2012) stating that responses to the graph necessitates a comprehension of the graph's relationship to the data context. While Kemp and Kissane (2010) primarily focused on teaching, their framework naturally guides students in solving a statistical problem, helping them navigate the steps effectively.

2.2.3 Statistical-Mathematical Knowledge

Knowledge of statistical concepts and related mathematical procedures is an obvious prerequisite to making sense of statistical information (Gal, 2002; Jureckova & Csachova, 2020). Statistical reports published in the media frequently use rational numbers, such as fractions, percentages and averages (Joram et al., 1995). Similarly, Gal and Geiger (2022) found that mathematical concepts, such as totals, percentages, proportions and rates, are often used in descriptive quantitative information to report phenomena related to COVID-19. These mathematical concepts are also employed in a number of persuasive ways, such as to influence public opinion and consumers (Weiland, 2019). Due to the importance of statistical-mathematical knowledge for students' SL, Watson and Callingham (2003) and Watson (2006) provided detailed descriptions of the levels of this knowledge required at each of their six hierarchical levels of SL.

However, a very salient question is: what kind of statistical understanding would be sufficient for students to be considered critical data consumers? Addressing that question, Gal (2002) suggested that students, as data consumers, need to know the functions of statistical concepts more than the underlying calculations. It could be risky to apply an algorithm without considering the context of using averages (Jacobbe & Carvalho, 2011; Landtblom, 2018). Students need to be aware of different averages, such as mean and mode, that are seemingly similar but can provide different interpretations and facts of the same dataset. One measure is preferable to another in specific situations (Groth & Bergner, 2006; Landtblom, 2018). Furthermore, they need to know how the mean is computed, the factors that could influence that computation (such as outliers and data distribution) and the conditions under which mean is applicable. It is also crucial to note that mode is not affected by extreme values, in contrast to the mean (Groth & Bergner, 2006; Landtblom, 2018). By understanding these concepts of averages, students can fully appreciate the meaning of statistical claims that use averages as a justification.

Regarding mathematical knowledge, students require an understanding of rational numbers to make sense of statistical information they encounter. When discussing the role of mathematical knowledge in SL, Gal (2002) mentioned the importance of numeracy skills and number sense. Numeracy skill involves understanding numbers, calculations, magnitudes and relationships (Hoffrén, 2021). Students must possess sufficient numeracy skills to accurately interpret numbers in statistical reports. Number sense is also essential for understanding diverse types of numbers (e.g., fractions, decimals and percentages). This mathematical knowledge is essential because the modern mass media environment requires young people to critically analyse statistical and mathematical data, evaluate news credibility and understand public policies (Gal & Geiger, 2022).

When dealing with graph-based statistical information, it is also necessary to have an adequate level of mathematical knowledge (Bursal & Yetiş, 2020; Ludewig, 2018). Ludewig et al. (2020) found that all basic numerical abilities are significantly associated with graph reading ability. Better performance in number line estimation predicted better graph interpretation (Ludewig et al., 2020). Students should be aware of the spatial location of specific value between the scale presented in the graph’s axis (Sharma, 2013a), especially when large numbers are displayed in thousands rather than in tens or hundreds. Incorrectly estimating the values to calculate or compare will lead to incorrect results and responses.

In conclusion, statistical-mathematical knowledge relates to the other two knowledge components—text and context, and representation. Altogether, these three knowledge components contribute to the students’ SL. Students require these three knowledge components to interpret data-based information, communicate relevant information using data, evaluate data-based claims and make decisions using relevant data-supported evidence. The following section discusses the existing SL assessment frameworks to gain insight on their coverage of the four SL skills and three SL components.

2.3 SL Assessments

This section aims to comprehensively explain the existing assessment frameworks for high school students' SL. In general, assessments are utilised in research for a variety of purposes, including informing instructors about student achievement and facilitating student learning (Sabbag et al., 2018). In statistics education, the ability to think and reason statistically could be one of the potential learning outcomes; however, assessing students' ability to think and reason statistically may be challenging (Woodard et al., 2020). Six SL assessment frameworks—developed by statistics educators who work with students across grade levels—were then identified and reviewed, focusing on the employed perspectives, constructs and levels. The purposes of these three focuses were to identify the perspective utilised by the previous frameworks, the skills evaluated under the constructs and the employed system of levels. The six assessment frameworks are the GAISE framework (Franklin et al., 2005; Franklin & Bargagliotti, 2020), the LOCUS framework (Whitaker et al., 2015), Mooney's framework (Mooney, 2002), Watson and Callingham's framework (Callingham & Watson, 2017; Watson, 1997; Watson, 2006; Watson & Callingham, 2003), the TIMSS (Mullis et al., 2012) and the PISA (OECD, 2014, 2023). The review of TIMSS and PISA focused on the data and chance domain and the uncertainty and data subscale respectively.

The review of the SL constructs employed in each of the six existing assessment frameworks identified two major perspectives: data producers and data consumers (see Table 2.1). The data producer perspective focuses on assessing students to think as young statisticians. This perspective is contained in the GAISE and LOCUS frameworks, which assess students using four constructs called problem-solving (Franklin et al., 2005; Whitaker et al., 2015). Through these constructs, students are confronted with a problematic situation requiring them to formulate a statistics question. They are then asked to collect data to address the statistical question they formulated, analyse the data they collected and eventually interpret the results they obtained.

Table 2.1

Six Existing Statistical Literacy (SL) Frameworks

SL perspective	Framework	Constructs	Hierarchical Levels	Participants
Data producers	The Guidelines for Assessment and Instruction in Statistics Education (GAISE; Franklin et al., 2005; Franklin & Bargagliotti, 2020)	Formulating questions, collecting data, analysing data and interpreting results	Levels A, B & C	Years 6–8
	The Levels of Conceptual Understanding in Statistics (LOCUS; Whitaker et al., 2015)	The same as GAISE	Levels A, B & C	Years 9–12

Table 2.1 (continued)

SL perspective	Framework	Constructs	Hierarchical Levels	Participants
Data consumers	Mooney's Framework (Mooney, 2002)	Describing, organising and reducing, representing, and analysing and interpreting data	Idiosyncratic, transitional, quantitative and analytical	Years 6–8
	Watson and Callingham's Framework (Callingham & Watson, 2017; Watson, 1997; Watson, 2006; Watson & Callingham, 2003)	Tier 1 (understanding basic statistical terms) Tier 2 (understanding basic statistical concepts in context) Tier 3 (using critical thinking)	Idiosyncratic, informal, inconsistent, consistent non-critical, critical and critical mathematical	Years 3–9 (in 2003) Years 5–11 (in 2017)
	The Trends in International Mathematics and Science Study (TIMSS; Mullis et al., 2012)	Knowing, applying and reasoning	Low, intermediate, high and advanced	Years 4 & 8
	The Programme for International Student Assessment (PISA; OECD, 2014, 2023)	Formulating, employing, and interpreting or evaluating	Levels 1–6 (in 2012 test) Levels 1c, 1b, 1a, and 2–6 (in 2022 test)	Students aged 15

In contrast, frameworks from the data consumer perspective assess students' abilities to respond to common data-driven information. Instead of producing data, students need to critically understand data. This perspective is shared across all four frameworks (Mooney, Watson and Callingham, TIMSS and PISA), each with distinct characteristics. The first of these frameworks is that of Mooney (2002), which establishes four constructs of SL from the data consumer perspective. In Mooney's constructs, students should demonstrate the ability to explain the information they obtain from a particular representation (describing data), the ability to group or order the data and describe the groups using measures of centre and spread (organising and reducing data), the ability to construct an alternative data display to communicate different ideas (representing data) and the ability to identify trends or make predictions from graphical representations (analysing and interpreting data). Rather than being hierarchical, these four constructs are sequential. In other words, the activity of 'describing data' comes before 'analysing and interpreting data' and both allow for four possible levels of responses.

The other three frameworks that adopt the data consumer perspective are Watson and Callingham’s (2003) framework, TIMSS (for its data and chance domain) and PISA (for its uncertainty and data subscale). Unlike Mooney’s framework, Watson and Callingham’s and TIMSS’ framework contain hierarchical constructs. However, PISA’s three constructs are not hierarchical; rather, each of them allows for responses across PISA’s hierarchical levels. The three constructs in TIMSS are called cognitive domains (Mullis et al., 2009), while the three constructs in PISA are called process categories (OECD, 2014, 2023). In TIMSS’ construct, knowing relates to students’ knowledge of statistical information and concepts in support of statements; applying assesses students’ ability to apply mathematical tools in a range of contexts; and reasoning measures students’ capacity for logical thinking and making generalisations. In terms of the uncertainty and data subscale, the three categories of PISA refer to formulating situations mathematically to represent contextual data information, employing a range of knowledge and skills to solve data-based problems and sensibly link information in a graph to textual information, and interpreting, applying and evaluating the information presented in graphs to present an argument or conclusion about contextual conditions.

Those differences in the SL perspectives and constructs contributed to each framework having a different system of levels. As shown in Table 2.1, the frameworks that employ the data producers' perspective have the same three-level hierarchical structure for assessing students' performance. In contrast, the categorisation of hierarchical levels in frameworks based on the data consumer perspective varies. The number of levels in frameworks based on the data consumer perspective ranges from four (in Mooney's framework and TIMSS) to six (in Watson and Callingham's framework and PISA).

Although they differ, a considerable effort was made, as part of this review, to align these systems of level. This alignment process makes it possible to compare the results of studies employing diverse level systems, as well as the SL of students whose levels was determined using diverse level systems. GAISE and LOCUS already use the same levels; therefore, alignment was unnecessary. In contrast, an alignment process was needed for the other four frameworks based on the data consumer perspective (PISA, TIMSS, Watson and Callingham, and Mooney). For example, although both the PISA framework and Watson and Callingham's framework have six hierarchical levels, this does not mean the hierarchies of those frameworks are parallel. Similarly, the four levels in the TIMSS framework and the four levels in Mooney's framework could not simply be paired. Additionally, aligning the levels enables the researcher to relate Mooney's four levels to Watson and Callingham's six levels, as these frameworks show similarities in their level names.

Table 2.2 demonstrates the process of aligning the PISA levels with Watson and Callingham's levels using their respective frameworks' level descriptors. A similar process was undertaken with the descriptors in the TIMSS framework and Mooney's framework. Regarding PISA framework, the descriptors for the uncertainty and data subscale in PISA 2022 was chosen because it is the current version of the PISA framework that assessed mathematics with the results published. However, descriptors for uncertainty were excluded.

This exclusion was necessary due to the present study focused on the data domain, not the uncertainty domain.

Table 2.2

Comparison between the Levels of the PISA Framework and Watson and Callingham's Framework

PISA 2022 (OECD, 2023)	Watson and Callingham (2003) and Callingham and Watson (2017)
Below level 1c No descriptors	*Idiosyncratic* Idiosyncratic engagement with context, tautological use of terminology and basic mathematical skills associated with one-to-one counting and reading cell values in tables.
Level 1c, 1b and 1a* Students at level 1a can typically read and extract data from charts or two-way tables and recognise how these data relate to the context. Students at level 1b can typically read information presented in a well-labelled table to locate and extract specific data values while ignoring distracting information.	*Informal* Only colloquial or informal engagement with a context that often reflects intuitive non-statistical beliefs, single elements of complex terminology and settings, and basic one-step straightforward table, graph and chance calculations. *Inconsistent* Selective engagement with the context, often in supportive formats, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas.

Table 2.2 (continued)

PISA 2022 (OECD, 2023)	Watson and Callingham (2003) and Callingham and Watson (2017)
Level 2 Students can identify, extract and comprehend statistical data presented in simple and familiar forms, such as a simple table, a bar graph or a pie chart. They can identify, understand and use basic descriptive statistical concepts in familiar contexts. At this level, students can interpret data in simple representations and apply suitable calculation procedures that connect given data to the problem context represented.	*Consistent non-critical* Appropriate but non-critical engagement with the context, multiple aspects of terminology usage, appreciation of variation in chance settings only, and statistical skills associated with the mean, simple probabilities and graph characteristics.
Level 3 Students can interpret and work with data and statistical information from a single representation that may include multiple data sources, such as a graph representing several variables, or from two related data representations, such as a simple data table and graph. They can work with and interpret descriptive statistical concepts and conventions in contexts and draw conclusions from data, such as calculations or using simple measures of centre ad spread. Students at this level can perform basic statistical reasoning in simple contexts.	
Level 4 Students can actively employ various data representations and statistical processes to interpret data, information and situations to solve problems. They can work effectively with constraints, such as statistical conditions that might apply in a sampling experiment. They can also interpret and actively translate between two related data representations (such as a graph and data table). Students at this level can perform statistical reasoning to make contextual conclusions.	
Level 5 Students can interpret and analyse a range of statistical data, information and situations to solve problems in complex contexts that require linking different problem components. They can use proportional reasoning effectively to link sample	*Critical* Critical, questioning engagement in familiar and unfamiliar contexts that do not involve proportional reasoning, but that do involve appropriate use of

Table 2.2 (continued)

PISA 2022 (OECD, 2023)	Watson and Callingham (2003) and Callingham and Watson (2017)
data to the population they represent. They can appropriately interpret data series over time. They are systemic in their use and exploration of data. Students at this level can use statistical concepts and knowledge to reflect, draw inferences and produce and communicate results.	terminology, qualitative interpretation of chance and appreciation of variation.
Level 6	*Critical mathematical*
Students can interpret, evaluate and critically reflect on a range of complex statistical data, information and situations to analyse problems. Students at this level bring insight and sustained reasoning across several problem elements; they understand the connections between data and the situations they represent and can use those connections to explore problem situations fully. They bring appropriate calculation techniques to bear to explore data, and they can produce and communicate conclusions, reasoning and explanations.	Critical, questioning engagement with context, using proportional reasoning, particularly in media or chance contexts, showing appreciation of the need for uncertainty in making predictions, and interpreting subtle aspects of language

Note. * there is no description for level 1c as there were no items used to assess this level in the PISA 2022 test.

The alignment process began by comparing the lowest to the highest level in each framework. The descriptor for PISA level 1c-1b-1a included students' ability to identify, read and locate specific data values from a simple well-labelled table and graph while recognising how these relate to the context. These descriptors did not match the descriptors of Watson and Callingham's idiosyncratic level, which covered students' idiosyncratic engagement with the context regardless of their ability to read cell values in tables. PISA level 1c-1b-1a captured more complex information about students' understanding of data than Watson and Callingham's idiosyncratic level. Consequently, Watson and Callingham's idiosyncratic level was aligned with PISA's 'below Level 1c'. PISA Level 1c-1b-1a was similar to Watson and Callingham's informal and inconsistent levels. The informal level of Watson and Callingham covered students' colloquial engagement with the context and their ability to perform basic, straightforward one-step table and graph calculations. Watson and Callingham's inconsistent level covered selective engagement with the context and appropriate recognition of conclusions. This alignment process continued until all levels in all frameworks had been compared. Table 2.3 presents the results of this alignment process across four different frameworks. However, it is important to note that the alignment does not provide a definitive comparison due to the complexities of each level system.

Table 2.3

Comparison SL Levels Between Four Frameworks

Watson and Callingham (2003)	Mooney (2002)	TIMSS (Mullis et al., 2012)	PISA (OECD, 2023)
Idiosyncratic	Idiosyncratic	Low	Below level 1c
Informal	Idiosyncratic	Low	Level 1c, 1b and 1a
Inconsistent	Transitional	Intermediate	Level 1c, 1b and 1a
Consistent non-critical	Quantitative	High	Levels 2, 3 & 4
Critical	Analytical	High and advanced	Level 5
Critical mathematical	Analytical	Advanced	Level 6

In summary, an SL assessment framework requires a perspective, a set of constructs and a set of levels. Assessing students' abilities to respond to data-based information portrays the perspective of data consumers. The frameworks involving this perspective have distinct constructs and different system of levels, making it challenging to compare the findings of various studies. To make it possible to compare the SL levels from research using various level systems, an alignment process was carried out. Finally, the result of reviewing the existing SL assessment frameworks ultimately serve as the basis for the development of a new assessment framework, as described in Section 2.6.

2.4 Differences in Students' SL

Developments in students' SL by grade level has been the topic of many SL studies. Some confirm growth in SL and graph skills with grade progression (Aoyama & Stephens, 2003; Bursal &Yetiş, 2020; Callingham & Watson, 2017), while others find no significant differences (Callingham & Watson, 2017; Yolcu, 2014). The non-significant difference in the students' SL between grades, particularly in students from adjacent grades, was claimed to be a result of the spiral curriculum (Yolcu, 2014). However, Indonesian students' SL levels across grade levels have not been widely studied in national journals, despite poor PISA and TIMSS performance.

Although many studies have been conducted to investigate the differences in students' SL according to grade level, limited studies have been conducted on how gender might affect SL. Moreover, some studies on gender differences (e.g., Carmichael & Hay, 2009; Chiesi & Primi, 2015) focused on students' interest in or attitudes towards statistics instead of their SL levels. Studies investigating the effect of gender on students' SL levels include Watson and Moritz (2000), Yolcu (2014), PISA 2003 (OECD, 2004) and PISA 2012 (OECD, 2014). Watson and Moritz conducted a study in Australia with students in Years 3 to 11, while Yolcu conducted a study in Turkey involving students in Years 6 to 8. In addition, PISA 2003 and 2012 both provided a broader picture of the gender differences in the uncertainty and data subscale. The PISA 2003 and 2012 reports covered students' levels on the uncertainty and data subscale among participating countries based on gender. These reports provided further insight into the trends that occurred over the decade from 2003 to 2012.

The findings from previous studies on the effect of gender on students' SL showed some consistency. In the uncertainty subscale of PISA 2003, gender differences were visible for 24 out of the 30 OECD countries (OECD, 2004). Boys outperformed girls in most countries. In Turkey, boys significantly outperformed girls, in contrast to Australia and

Indonesia, which were among the countries where there were no significant differences between boys and girls. In PISA 2012, the general trend was still consistent: boys outperformed girls on the uncertainty and data subscale across the participating countries (OECD, 2014). However, the trend changed for Australia and Turkey but remained the same for Indonesia. Specifically, there was a significant difference in Australia in favour of boys, and there was no significant difference between girls and boys in either Turkey or Indonesia. These findings indicate that Indonesian girls and boys performed equally poorly on this subscale over a decade, from 2003 to 2012.

Two other studies conducted in almost the same years as the PISA assessments were somewhat inconsistent with the PISA results. Watson and Moritz's (2000) findings in Australia were inconsistent with the PISA 2003 results and Yolcu's (2014) findings in Turkey were somewhat inconsistent with the PISA 2012 results. According to Watson and Moritz (2000), Year 9 girls performed significantly better than Year 9 boys. This result contradicted the PISA 2003 results reporting that Australian boys outperformed Australian girls. Similarly, according to Yolcu (2014), in Turkey, there was a significant difference in terms of gender in favour of girls. This finding contradicted the PISA 2012 results for Turkey students, which showed no significant difference. Such contradictions might have been caused by the participants and the instruments. For instance, Yolcu's study involved younger students and a smaller number of students than PISA. In addition, Yolcu's problems were derived from those of Watson and Callingham (2003), which, to a certain extent, differed from the problems tested by PISA.

In addition to grade and gender differences, school characteristics and region are considered to influence students' performance. However, the existing studies on SL have not focused on this aspect worldwide, including Indonesia that has a unique school system (its characteristics is discussed in Chapter 3). In Indonesia, there is a commonly held belief that students in private schools perform worse than those from public schools (Bedi & Garg, 2000; Muttaqin et al., 2020). Although there are private schools with high achievement levels, their number is small (Bedi & Garg, 2000). Similarly, in Indonesia, students from schools under the Ministry of Religious Affairs (MoRA) are commonly considered to show lower performance than those from schools under the Ministry of Education, Culture, Research and Technology (MoEC-RT; Newhouse & Beegle, 2006). This is supported by the fact that students in schools under MoRA have more subjects to learn than those in schools under MoEC-RT—sometimes almost double. As a result, they have less time to learn mathematics and, moreover, statistics at schools. Finally, schools in Java and Bali are significantly more developed than those in the eastern parts of Indonesia (Azzizah, 2015). This also applies to schools in urban areas, which have more facilities than schools in rural areas. Hence, the region where the school is located might influence students' performance.

Since data on the Indonesian students' performance in statistics are unavailable, their performance in UN (Indonesian national examination) for mathematics is used instead. Table 2.4 compares the students' average scores in UN based on school status—state in comparison to private school (Pusat Penilaian Pendidikan Kemdikbud, 2023). For Year 9 students, the national average on the students' performance was nearly the same, with the exception of 2018 showing students from state school got higher score than those from private schools. In comparison, Year 12 students at state schools tend to have higher average over the years. This reinforced the commonly belief that students from state schools are 'better' than private school students. However, the case for statistical domain has not been well examined to determine whether there are disparities in students' SL based on school status. This is important to ensure that students in both state and private schools have the same opportunity to be statistically literate.

Table 2.4

National Average of Year 9 and Year 12 Students' Mathematics Score from State and Private Schools

	2016	2017	2018
Year 9			
State	50.14	50.38	44.16
Private	50.43	50.19	41.91
Year 12 (Science)			
State	53.69	41.98	37.35
Private	51.36	39.91	34.29

Similarly, data on the students' performance in mathematics is used to compare the schools under MoEC-RT and MoRA. Table 2.5 compares the students' average scores in UN based on school type (Pusat Penilaian Pendidikan Kemdikbud, 2023). For Year 9 students, the national average score on the students' performance was nearly the same in 2017—between schools under MoEC-RT and MoRA. However, in 2016 schools under MoRA had higher average in contrast to 2018 showing students from MoEC-RT have higher average score. This is not the case for Year 12 students, in which students from MoEC-RT tend to have higher average over the years. This case for Year 12 students reinforced the commonly belief that schools under MoEC-RT are 'better' than schools under MoRA. However, the case for statistical domain has not been well examined to determine whether there are disparities in students' SL based on school type. Again, this is important to ensure that students from both MoEC-RT and MoRA schools have the same opportunity to be statistically literate.

Table 2.5

National Average of Year 9 and Year 12 Students' Mathematics Score from MoEC-RT and MoRA Schools

	2016	2017	2018
Year 9			
MoEC-RT	49.84	50.34	44.05
MoRA	51.80	50.36	41.16
Year 12 (Science)			
MoEC-RT	53.54	41.92	37.25
MoRA	50.16	38.55	32.40

Similar to the school status and school type, regional disparity can be seen by comparing the average score in national, provincial and regional levels. In national level, data shows that the average score of students from Western parts of Indonesia are higher than Eastern parts of Indonesia (Pusat Penilaian Pendidikan Kemdikbud, 2023). Furthermore, within the same province, the schools in big cities show higher average score than schools in small cities. This trend continues to the regional levels, in which schools in the urban areas tend to have higher average score than schools in rural areas. However, there is a little attention to regional disparities in students' SL, despite the fact that the differences are quite pronounced. In order to determine if students from different regions have equal opportunities to become statistically literate, this study also explores regional disparities in students' SL.

2.5 Students' Challenges and Understandings in SL

The underperformance of more than 70% of Indonesian students in the uncertainty and data subscale of PISA clearly indicates that these students had challenges solving data-based problems (OECD, 2004, 2014). However, PISA results only capture the general trend and do not specifically present the causes and types of student errors (Sari & Valentino, 2017). From the perspective of a teacher, understanding the cause of students' errors or challenges can serve as a preliminary step towards improving students' performance (Case & Jacobbe, 2018; Wijaya et al., 2019). From the perspective of the students themselves, knowing their errors can help them overcome their own challenges (Sharma, 2006). Jureckova and Csachova (2020) identify factors like context, statistical and mathematical knowledge behind the challenges students encountered. Thus, revealing students' challenges in responding to the data-based items could help shed light on the causes of students' errors. Further, identifying challenges students encountered is a crucial step in assessing students' reasoning (Muttaqin et al., 2017).

While the use of context in mathematics problems can promote various opportunities for students to learn mathematics, it is also considered a factor causing challenges students encounter (Parmar & Signer, 2005; Wijaya et al., 2014). In schools, the use of contextual problems in statistics, as part of mathematics, provides students with the opportunity to make sense of data in a context. However, they may have challenges in solving the data-based problems when they lack context knowledge (Jureckova & Csachova, 2020; Sharma, 2006; Sharma et al., 2011; Sharma, 2013b). In PISA, TIMSS and other SL-assessing studies, context appears in the form of textual information and normally appears in paragraphs. The length of the text and the use of certain and technical terminology may trigger students' challenges. For example, students may misinterpret the word *summarise* in the question and fail to translate it into statistical meaning (Sharma, 2006). Students may also misinterpret the term *taller than* as *the tallest* (Parmar & Signer, 2005), *at least* as *less than* and *at most* as *more than* (Sharma, 2018a). Consequently, their lack of text and context knowledge causes them to only relate the problem's context to their personal experiences (Wijaya et al., 2014) and prevents them from having a critical interpretation.

In addition to the context, the data representation could be another source of challenges students encounter. A representation contains conventions that students need to understand. According to Schield (2000), tables' titles, subtitles, and column and row headings can be a source of students' challenges. As a result of these challenges, students find the mode, or other averages more challenging to understand (Landtblom, 2018; Leavy & O'Loughlin, 2006). Similarly, graph features, such as colour, size, scale and legends, are a source of students' challenges in interpreting information from graphs (Glazer, 2011). Sharma (2006) found that although students could read tables and graphs, they were unsuccessful in drawing inferences, in solving problems if explicit information was unavailable and in solving data-based problems that required higher order thinking.

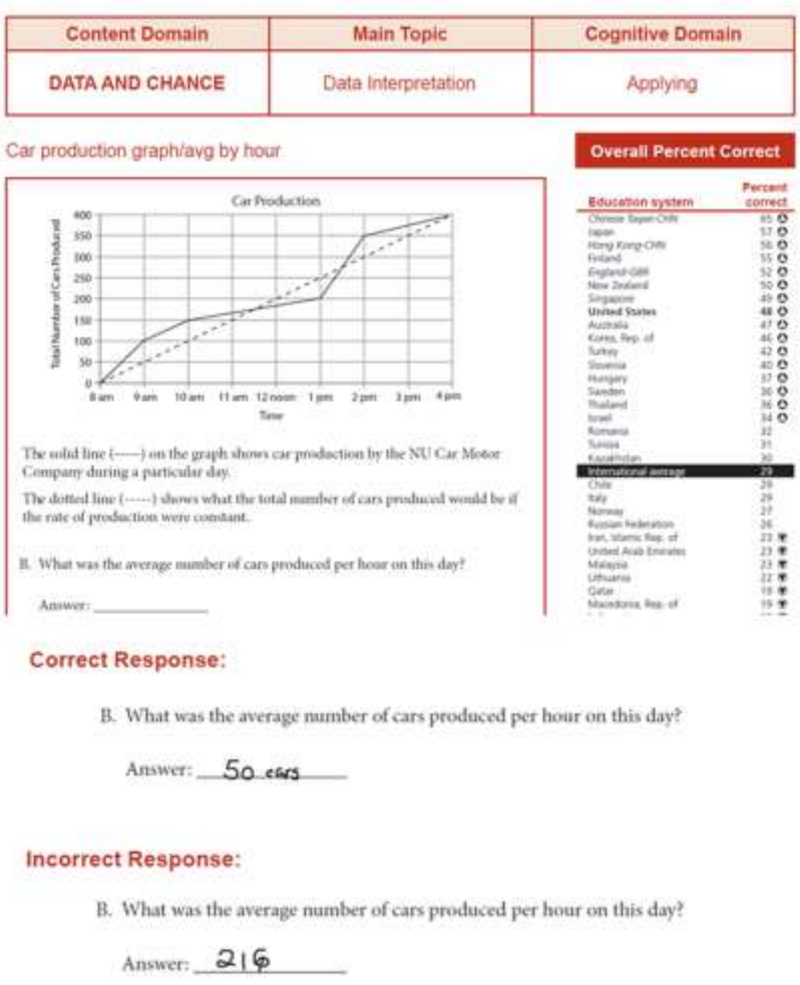
Students' challenges and inability to understand representations also attributable to their statistical-mathematical knowledge. When a representation contains data that must be interpreted and processed using their statistical-mathematical knowledge, they may encounter challenges. For instance, students struggle to understand a table or graph consisting of rates and percentages that must be compared using an arithmetic comparison (Schield, 2000). The content of a graph (such as missing data, scale, or patterns) and a lack of prior statistical-mathematical knowledge are additional sources of students' challenges (Glazer, 2011; Sharma, 2006). Students may know the arithmetic procedure to calculate the mean; however, they may not have a conceptual understanding of the mean (Batanero et al., 1994). Further, they are frequently confused by the procedures for finding mean and median and are unable to select the most appropriate measure of central tendency when responding to data-based problems (Zawojewski & Shaughnessy, 2000).

Figure 2.2 shows one data-based item from TIMSS with two responses from students. This question asks students to determine the average number of cars produced per hour from the data presented in a line graph. The international average percentage of correct answers was 29%. Taiwanese students attained the highest percentage of correct answers (65%), followed by students from Japan, Hong Kong, Finland, England and New Zealand (50–60%).

In contrast, only 19% of students in Indonesia were successful at solving this problem. In other words, 81% of Indonesian students found this problem too difficult to solve. However, there is not enough data to clarify what caused 19% of Indonesian students to answer correctly and 81% to answer incorrectly. There must be a difference in students' thought processes causing them to answer correctly or incorrectly, and this needs further investigation. Therefore, written or spoken responses are crucial to clarifying whether the context, the line graph, the term *average* or other parts of the problems caused students to succeed or fail in answering this question.

Figure 2.2

TIMSS Mathematics Item and Examples of Student Challenges and Understanding



Content Domain	Main Topic	Cognitive Domain
DATA AND CHANCE	Data Interpretation	Applying

Car production graph/avg by hour

The solid line (——) on the graph shows car production by the NU Car Motor Company during a particular day.

The dotted line (-----) shows what the total number of cars produced would be if the rate of production were constant.

B. What was the average number of cars produced per hour on this day?

Answer: ____________

Overall Percent Correct

Education system	Percent correct
Chinese Taipei-CHN	65 ▲
Japan	57 ▲
Hong Kong-CHN	56 ▲
Finland	55 ▲
England-GBR	52 ▲
New Zealand	50 ▲
Singapore	49 ▲
United States	**48** ▲
Australia	47 ▲
Korea, Rep. of	46 ▲
Turkey	42 ▲
Slovenia	40 ▲
Hungary	37 ▲
Sweden	36 ▲
Thailand	36 ▲
Israel	34 ▲
Romania	32
Tunisia	31
Kazakhstan	30
International average	29
Chile	29
Italy	29
Norway	27
Russian Federation	26
Iran, Islamic Rep. of	23 ▼
United Arab Emirates	23 ▼
Malaysia	23 ▼
Lithuania	22 ▼
Qatar	19 ▼
Macedonia, Rep. of	19 ▼

Correct Response:

B. What was the average number of cars produced per hour on this day?

Answer: 50 cars

Incorrect Response:

B. What was the average number of cars produced per hour on this day?

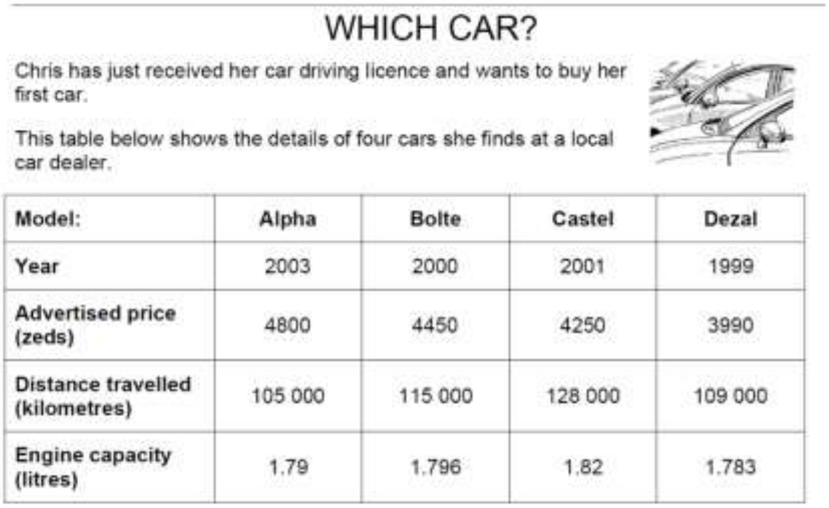
Answer: 216

Note. Car Production is a TIMSS released item (TIMSS & PIRLS, 2011)

Figure 2.3 presents another example of data-based items, obtained from the PISA 2012 reports; it includes two questions from a single context. The first question was categorised under ‘uncertainty and data’, while the second question was categorised under ‘quantity’. Additionally, the first question was designed to help measure whether students can solve this ‘below Level 1’ problem, while the second question was designed to help determine whether students were at PISA Level 4 (OECD, 2014, p. 60). The first question requires students to have basic knowledge of the row and column conventions of a table to identify when the three conditions are all met. The solution also requires basic understanding of large whole numbers, but this knowledge is unlikely to be the main source of challenges for students aged 15 years old. In contrast, the second question is expected to be a much larger source of cognitive demand than identifying the correct data from the table. Students’ challenges in dealing with decimal numbers and percentages are reflected in the empirical results. Further examination is required to reveal the source of students’ challenges, particularly whether these challenges arise from the textual information, the data presented in the table or the need to perform mathematical operations using percentages.

Figure 2.3

PISA Mathematics Item and Scoring Guides



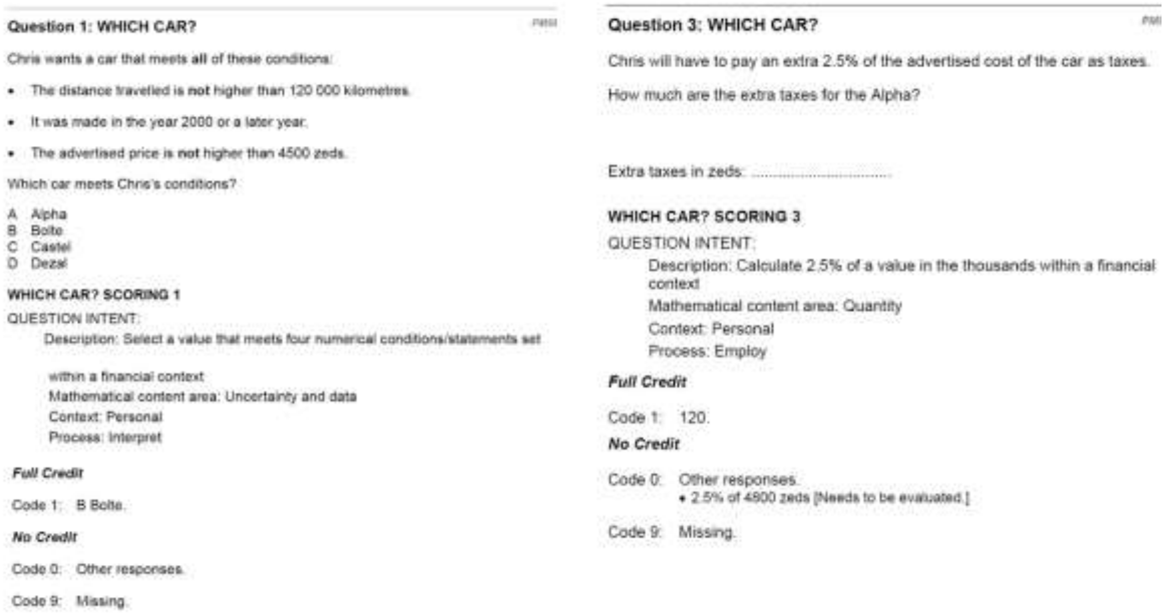
WHICH CAR?

Chris has just received her car driving licence and wants to buy her first car.

This table below shows the details of four cars she finds at a local car dealer.

Model:	Alpha	Bolte	Castel	Dezal
Year	2003	2000	2001	1999
Advertised price (zeds)	4800	4450	4250	3990
Distance travelled (kilometres)	105 000	115 000	128 000	109 000
Engine capacity (litres)	1.79	1.796	1.82	1.783

Figure 2.3 (continued)



Question 1: WHICH CAR?

Chris wants a car that meets **all** of these conditions:

- The distance travelled is **not** higher than 120 000 kilometres.
- It was made in the year 2000 or a later year.
- The advertised price is **not** higher than 4500 zeds.

Which car meets Chris's conditions?

A Alpha
B Bolte
C Castel
D Dezal

WHICH CAR? SCORING 1

QUESTION INTENT:

Description: Select a value that meets four numerical conditions/statements set within a financial context
Mathematical content area: Uncertainty and data
Context: Personal
Process: Interpret

Full Credit

Code 1: B Bolte.

No Credit

Code 0: Other responses.

Code 9: Missing.

Question 3: WHICH CAR?

Chris will have to pay an extra 2.5% of the advertised cost of the car as taxes.

How much are the extra taxes for the Alpha?

Extra taxes in zeds: ....

WHICH CAR? SCORING 3

QUESTION INTENT:

Description: Calculate 2.5% of a value in the thousands within a financial context
Mathematical content area: Quantity
Context: Personal
Process: Employ

Full Credit

Code 1: 120.

No Credit

Code 0: Other responses.
- 2.5% of 4800 zeds [Needs to be evaluated.]

Code 9: Missing.

Note. Which Car? is a PISA released item (OECD, 2013b)

In conclusion, it is necessary to determine the root reasons of the challenges Indonesian students had when attempting to solve data-based problems. Making this identification can be the first step towards enhancing their SL and reasoning. To achieve these objectives, a new assessment SL framework was proposed, as explained in subsequent section.

2.6 A New Assessment Framework

This section discusses the assessment framework proposed in this study. Building upon the insights from the previous section's review, an assessment framework to characterise students’ SL requires three aspects: a perspective, a set of constructs and a set of levels. Given that all high school students, as citizens, are data consumers, a new SL framework was designed to adopt the perspective of data consumers rather than that of data producers. The four skills needed by students to respond to data-based information became the SL constructs. Rather of being hierarchical, these four skills are sequential, which suggests that one skill is deemed neither simpler nor more sophisticated than the other skills. Additionally, each SL skill necessitates three knowledge components, which are established as sub-constructs. Finally, the framework adapts and refines the existing levels to align with the constructs and sub-constructs. Table 2.6 summarises the characteristics of the four skills.

Table 2.6

Four Skills Required by Students to Effectively Respond to Statistical Information

Skills	Characteristics
Interpreting	The capability to derive qualitative meaning from quantitative data
Communicating	The capability to effectively share or discuss statistical information with others by selecting the most significant data
Evaluating	The capability to argue statistical claims or arguments with reasonable and critical evidence
Decision-making	The capability to make informed decisions based on statistical arguments

Among the existing levels from the six assessment frameworks, the SL levels used by Watson and Callingham (2003) and Callingham and Watson (2017) are the most appropriate levels for characterising students' SL in this study. These six hierarchical levels have been empirically studied extensively over a decade to describe the development of a critical SL in students (Weiland & Sundrani, 2022), confirmed to be relevant over two decades (Callingham & Watson, 2017) and employed in many SL assessment studies (Koga, 2022a). More importantly, these levels are appropriate because they are based on descriptors that can be related to the three SL components (text and context, representation and statistical-mathematical knowledge). In increasing order, the levels are as follows: idiosyncratic, informal, inconsistent, consistent non-critical, critical and critical mathematical.

Further, the original descriptors of the six hierarchical levels needed to be clearly classified under the three SL components to provide a scoring guide. This classification is supported by other studies that described the three components at each of the six hierarchical levels (e.g., Sharma et al., 2012; Sharma, 2013a; Watson, 2006; Watson & Callingham, 2003). Table 2.7 presents the original and modified descriptors of the six levels. As the modified descriptors show, the six hierarchical levels can be characterised using keywords. At the idiosyncratic level, students use personal and intuitive viewpoints. At the informal level, students use colloquial or daily related viewpoints. At the inconsistent level, students apply content knowledge inappropriately or without statistical reasoning. According to Sharma et al. (2012), these first three levels indicate the non-statistical thinking of students, and in this study, they are called the lower group level. Additionally, at the consistent non-critical level, students are likely to be successful in solving problems but provide uncritical responses. At the critical and critical mathematical levels, students can think critically, but the complexity of their thinking varies. In this study, they are called the upper group level expressing levels of students using statistical thinking when responding to statistical information.

Table 2.7

Characteristics of the Six Hierarchical Levels

Level	Original descriptor	Modified descriptor
Idiosyncratic	Idiosyncratic engagement with the context, tautological use of terminology and basic mathematical skills associated with one-to-one counting and reading cell values in tables.	This level indicates that a student’s response is dominated by personal belief and experience. Students will use the provided context in a straightforward manner. When examining data representations, they can read specific values from simple graphs or tables. In terms of statistical-mathematical knowledge, students can perform simple calculations and one-to-one counting from data values on graphs or tables.

Table 2.7 (continued)

Level	Original descriptor	Modified descriptor
Informal	Only colloquial or informal engagement with the context, often reflecting intuitive non-statistical beliefs, single elements of complex terminology and settings, and basic one-step straightforward table, graph and chance calculations.	Response at this level demands engagement with the context to a greater extent than at the first level, although the response may still be intuitive, non-statistical or focused on irrelevant aspects of the task context. When reading a table or graph, students may be successful at some of the more basic table and graph reading tasks, such as comparing cells to determine the highest or most frequent value and identifying the smaller data value. This can result in an informal arithmetic calculation of a single idea, such as an average.
Inconsistent	Selective engagement with the context, often in supportive formats, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas.	Response at this level requires more engagement with the context than in the previous two levels, but this is dependent to some extent on the format of the question, which may provide some support. The statistical ideas required at this level are qualitative rather than quantitative, and appropriate conclusions may not be accompanied by suitable justifications (inappropriate explanations). For data representation, this level demands one summary statement when interpreting a basic unlabelled graph or a labelled graph with no association when an association is intended (fail to show relationship).
Consistent non-critical	Appropriate but non-critical engagement with context, multiple aspects of terminology usage, appreciation of variation in chance settings only, and statistical skills associated with the mean, simple probabilities and graph characteristics.	At this level, the responses demand a consolidation of appropriate contextual but non-critical engagement by students in various contexts. Graph recognition responses demand single, or at least partially correct, comparison of data in a table or graph and recognition of the highest data value and the range of the data. Accurate use of statistical skills associated with simple statistics and graph characteristics is required at this level. The mathematical and statistical skills required include those associated with the mean and graph characteristics in straightforward settings.

Table 2.7 (continued)

Level	Original descriptor	Modified descriptor
Critical	Critical, questioning engagement in familiar and unfamiliar contexts that do not involve proportional reasoning but involve appropriate terminology, qualitative interpretation of chance and appreciation of variation.	Sophisticated use of proportional reasoning is not required, but critical thinking is expected in certain contexts, particularly familiar ones. In terms of graphical competence, responses require the creation of a summary based on data in the table and graphs. Critical use of statistical skills associated with non-simple statistics and table, or graph characteristics is required for statistical and mathematical knowledge.
Critical mathematical	Critical, questioning engagement with the context using proportional reasoning, particularly in media or chance contexts, an appreciation of the need for uncertainty in making predictions and interpreting subtle aspects of language.	Proportional reasoning skills are often required at this highest level, particularly those that show critical engagement with the context. In terms of graphical competence, an appropriate summary statement involving the context is required, rather than just data reading. Sophisticated statistical and mathematical knowledge is evidenced by a knowledge of ratio and part–whole relationships.

Note. The original descriptor is taken from Watson and Callingham (2003) and Callingham and Watson (2017).

For each of the six hierarchical levels, from the idiosyncratic level to the critical mathematical level, descriptors that capture students' understanding of each component were established. These descriptors are referred to as component-based descriptors. They were used as a guide to assign the most appropriate level to each component. Based on the modified descriptors in Table 2.7, the characteristics of each level in relation to the three SL components were further examined. For example, the characteristics of Level 1 are personal engagement with context, basic graph reading and one-to-one counting or simple calculation. These characteristics were classified under the three components of SL. First, personal engagement with the context characterises the text and context component. Second, basic graph reading indicates the required representation component. Finally, basic mathematical skills associated with one-to-one counting characterise the statistical-mathematical knowledge component.

Similarly, the characteristics of Level 6, the critical mathematical level, in relation to the three components of SL are questioning of contexts, critical summarising of the association of the variables shown in a graph or table and displaying statistical and mathematical skills. This Level 6 characteristic was then classified under the three components. Critical and questioning engagement with familiar and unfamiliar contexts relates to the text and context component. To demonstrate the representation component of their SL at this level, students must critically summarise the associations between variables shown in a graph or table and relate them to its context. In relation to the statistical-mathematical knowledge component, students must display sophisticated or critical statistical and mathematical skills associated with statistical concepts such as central tendency and dispersion measures. Table 2.8 presents the component-based descriptors for all six levels.

Table 2.8

Component-Based Descriptors

Level	Component-based descriptors
Idiosyncratic	*Text and context*: Students demonstrate non-existent or personal engagement with contexts. *Representation*: Students show that their personal beliefs and experiences underlie their basic graph and table reading (e.g., reading cell values). *Statistical-mathematical*: Students guess the answer, perform one-to-one counting, pick a random value, perform basic calculations, select the largest number or take other unreasonable steps.

Table 2.8 (continued)

Level	Component-based descriptors
Informal	*Text and context*: Students show engagement with contexts, but their engagement is colloquial or informal (reflecting intuitive or non-statistical beliefs) and reflects irrelevant aspects of the context. *Representation*: Students may be successful at some of the more basic table or graph reading, such as comparing cells to determine the highest or most frequent value and identifying the smaller data value. *Statistical-mathematical*: Students perform basic one-step table and graph calculations (such as addition and subtraction) based on the values observed, but sometimes accompany the calculation with an imaginative story.
Inconsistent	*Text and context*: Students demonstrate selective or inconsistent engagement with contexts (depending, to some extent, on the format of the items). *Representation*: Students tend to interpret the graphical or tabular details rather than the context of the graph or table and fail to describe the relationship between data. *Statistical-mathematical*: Students make conclusions, but those conclusions are not always accompanied by suitable statistical or mathematical justifications.
Consistent non-critical	*Text and context*: Students show appropriate engagement with contexts, but often do so in a non-critical manner. *Representation*: Students make sense of the data presented in a graph or table while partially recognising the context, focus on a single relevant aspect of the data or compare the data within the table or graph. *Statistical-mathematical*: Students accurately and appropriately use simple statistical and mathematical concepts, including those associated with graph characteristics.
Critical	*Text and context*: Students demonstrate critical engagement with familiar contexts and less-critical engagement with unfamiliar contexts. *Representation*: Students demonstrate awareness of relevant features of a graph or table and awareness of the integration of more than one relevant aspect of data to show a relationship. *Statistical-mathematical*: Students demonstrate qualitative interpretation and sophisticated use of mathematical or statistical concepts.

Table 2.8 (continued)

Level	Component-based descriptors
Critical mathematical	*Text and context*: Students demonstrate critical, questioning engagement with familiar and unfamiliar contexts. *Representation*: Students critically summarise the association between the variables shown in a graph or table and relate it to the context. *Statistical-mathematical*: Students perform sophisticated or critical statistical and mathematical tasks associated with statistical concepts such as central tendency and dispersion measures.

These component-based descriptors complement the proposed assessment framework to characterise students' SL level. In other words, these descriptors capture the spectrum of students' responses as data consumers when they are asked to respond using one of the four SL skills and involving all the three SL components. Additionally, this framework provided the foundation for instrument development and piloting (described in Chapter 4) as well as the implementation studies (explained in Chapter 5). The four SL skills and three SL components were used as the basis for this study's item development and interview protocol, while the descriptors were used for scoring guide.

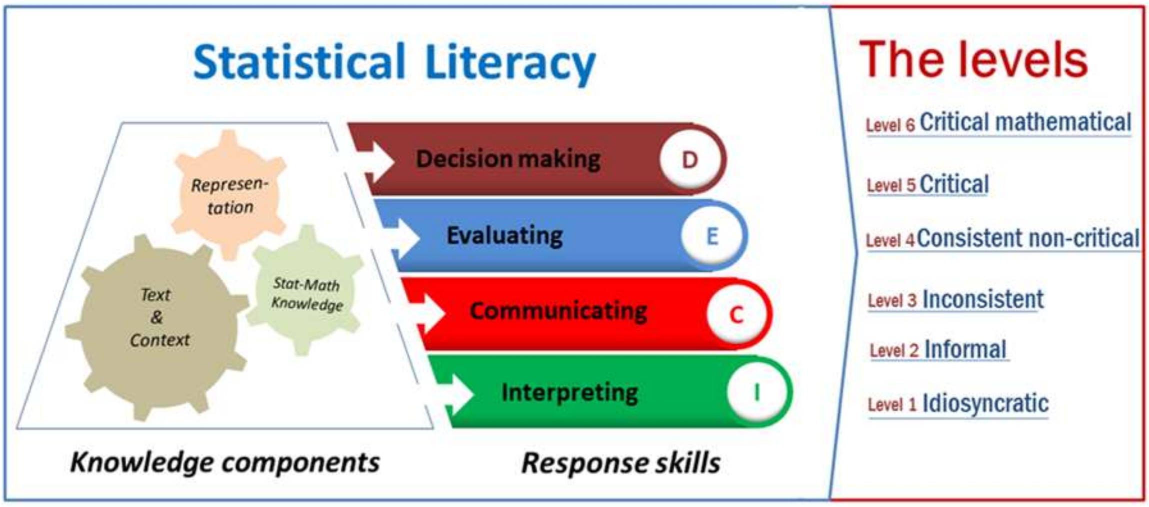
2.7 Chapter Summary

This chapter has reviewed the definition of SL and the assessment frameworks that have been used to assess high school students' SL. According to the literature, critical response to quantitative information is a strong indicator of students' SL and involves four complex response skills: data interpretation, data-based communication, data evaluation and data-driven decision-making. In addition, students' critical responses are strongly influenced by their appreciation of so-called knowledge components (text and context, representation and statistical-mathematical knowledge). Therefore, assessing these three SL components can provide a great deal of information about students' SL in relation to the four complex response skills listed above. In addition, assessing these three SL components can reveal students' challenges and understandings when they solve data-based problems.

As a result, an alternative SL assessment framework consisting of four skills and three components (see Figure 2.4) has been proposed in this chapter to assess Indonesian high school students' SL. A hierarchy of six levels was established, with descriptors for each component. The levels, from lowest to highest, are as follows: idiosyncratic, informal, inconsistent, consistent non-critical, critical and critical mathematical.

Figure 2.4

Theoretical Framework for Assessing SL



Note. This illustration of SL assessment framework was taken from Kurnia et al., (2023)

The following chapter, Chapter 3, was then used to identify opportunities to conduct research with Indonesian high school students using this assessment framework. Consequently, Chapter 3 examines the Indonesian curriculum, mathematics textbooks, teachers' demography and pedagogy, and international and national assessments using this framework as a review lens. Following on, the framework was then used to guide instrument development (as detailed in Chapter 4) and applied for data analysis to determine the SL levels of students from distinct cohorts (as explored in Chapter 5).

Chapter 3: Study Context

Assessing SL for Indonesian high school students requires a sufficient understanding of their contextual background. Obtaining such an understanding requires a review of the extent to which the Indonesian mathematics curriculum provides students with opportunities to be statistically literate. Indonesian education relies heavily on the use of textbooks, and the textbook defines what students learn (Büscher, 2022b; Landtblom, 2018; Ponte & Marques, 2011). Therefore, this chapter also reviews the extent to which Indonesian mathematics textbooks for high schools provide activities and tasks that can support students to be critical of data-based information. Subsequently, the teachers' demography and pedagogy were then reviewed to reveal their qualification. Further, the UN (Indonesian national examination) and international mathematics assessments in which Indonesian students have participated were reviewed to reveal the typical items used to assess these students' understanding of statistics.

Using the SL assessment framework proposed in the previous chapter, the reviews were conducted to identify opportunities to assess the SL of Indonesian high school students. This meant the review utilised that framework's constructs (the four SL skills) and sub-constructs (the three SL components). Section 3.1 explains the Indonesian education system in general and the high school statistics curriculum to provide the context of this study. Further, Section 3.2 describes to what extent the Indonesian curriculum provides opportunities for students to be statistically literate, even though the performance of Indonesian students in PISA data-based items has been low during the last two decades. Section 3.3 reviews the statistical content of Indonesian mathematics textbooks and Section 3.4 reviews the demography and pedagogy of Indonesian mathematics teachers. Section 3.5 reviews statistical items in both national and international assessments. Finally, Section 3.6 concludes this chapter with a summary.

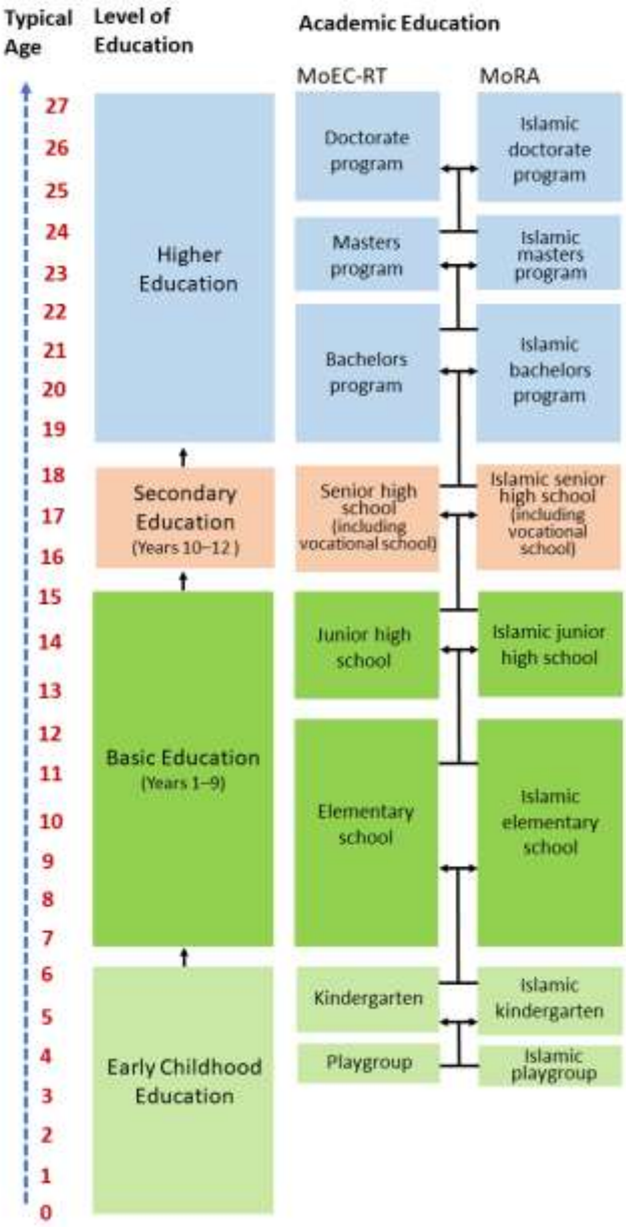
3.1 Indonesia’s Education System

Indonesia is a developing country in South-East Asia with 38 provinces (in 2024) across thousands of islands. It is the fourth most populated country worldwide, with more than 275 million people (Badan Pusat Statistik, 2023), of whom more than 44 million are of school age (Kemdikbud, 2023). Indonesia faces numerous challenges in educating its citizens to participate in society (Kemdikbud, 2013). To improve its quality of education, Indonesia has changed its curriculum many times since Independence Day (Patahuddin et al., 2018), and has participated in international tests such as PISA and TIMSS. It is recorded that its curriculum has changed four times in the last two decades (Inspektorat Jenderal Kemdikbud, 2023) with schools using, in turn, the Competence-Based Curriculum (2004), the School-Based Curriculum (2006), Kurikulum 2013 abbreviated K13 (2013) and Kurikulum Merdeka [Emancipated Curriculum] (2022). The curriculum change from School-Based Curriculum to K13 was partly influenced by the underperformance of Indonesian students in the international tests since their first participation (Fitriyah, 2020; Pratiwi, 2019; Zulkardi & Putri, 2019).

Indonesia’s education system is the fourth largest in the world (Patahuddin et al., 2018) and is administered by MoEC-RT and MoRA. Figure 3.1 maps the education system in Indonesia. Each education level, from early childhood to university, is organised by either MoEC-RT or MoRA. Additionally, schools under MoEC-RT and MoRA have either private or public status. The compulsory length of education in Indonesia is 12 years: six years of elementary school, three years of junior high school and three years of senior high school. In other words, students finish their schooling at around 18 years old. They can proceed directly to university; if they do not, they may still enter a state-owned university at any time within four years. For a private university, there is no age restriction.

Figure 3.1

Indonesian Education System


Typical Age
Level of Education
Academic Education
MoEC-RT
MoRA
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Higher Education
Secondary Education (Years 10–12)
Basic Education (Years 1–9)
Early Childhood Education
Doctorate program
Masters program
Bachelors program
Senior high school (including vocational school)
Junior high school
Elementary school
Kindergarten
Playgroup
Islamic doctorate program
Islamic masters program
Islamic bachelors program
Islamic senior high school (including vocational school)
Islamic junior high school
Islamic elementary school
Islamic kindergarten
Islamic playgroup

Note. This education system is adapted from Jupri (2015, p. 20) and Kemdikbud (2016, p. 14-15)

In terms of school status, a significant number of the private schools in Indonesia have emerged as major providers of education services in comparison to their public counterparts. Prior to 2015, more than half of the junior high school (Years 7–9 or approximately 13–15 years of age) and almost 70% of the senior high school (Years 10–12 or approximately 16–18 years of age) in Indonesia were private schools (Bappenas, 2015). In 2021, private schools represented 41.7% of junior high schools and 50.24 % of senior high schools, a decline from the previous years (Badan Pusat Statistik, 2022; Kemdikbud, 2023). However, only 26.96% of junior high school students and 26.16% of senior high school students attended private schools in 2022. Given that there are over 10 million students in junior high schools and over 5 million in senior high schools (Badan Pusat Statistik, 2022), the number of students attending private schools is still enormous. Moreover, graduates of private schools perform better in the labour market, despite the widely held belief that public schools are 'better' than private schools (Bedi & Garg, 2000).

All students were required to take the UN in Year 9 and Year 12 (the end of schooling). However, this examination ended in 2019 and it was replaced by the *Asesmen Nasional* (Computer-Based National Assessment) in 2021 (The Regulation of the Minister of National Education, Culture, Research and Technology, 2021; Kharismawati, 2022). In the UN, mathematics was one of the subjects tested. In the junior high school (Year 9) UN, the other subjects were Bahasa Indonesia, English and science. In the senior high school (Year 12) UN, the other subjects were Bahasa Indonesia, English and three additional subjects based on the student's intended major. There are at least three majors that students in senior high school can choose: Science, Social studies and Language. In the UN, the additional subjects for Year 12 Science students were physics, chemistry and biology; for Year 12 Social studies students, they were economics, sociology and geography; for Year 12 Language students, they were anthropology, Indonesian literature and one other language (Arabic, French, German, Japanese or Mandarin).

Statistics is frequently viewed as part of mathematics curriculum (Büscher, 2022b), including in Indonesian curriculum. As statistics is taught as part of mathematics, students' understanding of statistics was also assessed in the UN. There are four domains that were assessed in the Year 9 UN for mathematics examination: number, algebra, geometry and measurement, and statistics and probability. In the Year 12 UN, the mathematics domains tested varied depending on the student's major. Table 3.1 presents the mathematics domains tested in the Year 12 UN from 2016 to 2019. Despite the apparent differences between the Science, Social studies and Language majors, statistics and probability were tested for students undertaking all three. Although the term 'statistical literacy' was often referred to as 'statistics and probability' (and probability is called 'uncertainty' in PISA and 'chance' in TIMSS), the present study specifically focused on the statistics domain.

Table 3.1

Mathematics Domains Tested for Year 12 Students in UN (Indonesian National Examination)

	Student's major		
	Science	Social studies	Language
Mathematics Domains	Algebra	Algebra	Algebra
	Calculus	Calculus	–
	Geometry and Trigonometry	Geometry and Trigonometry	Geometry and Trigonometry
	Statistics and Probability	Statistics and Probability	Statistics and Probability

Note. The coverage of these domains in Year 12 UN is derived from Pusat Penilaian Pendidikan Kemdikbud (2023).

3.2 Curriculum Review

Due to the continuous refinements that have been made to the Indonesian curriculum, it is important to note that the curriculum review presented here was conducted during the period when this study's data were collected, 2019, meaning the curriculum that was reviewed was K13. The review showed that K13 created reasonable opportunities to support students to be statistically literate. The curriculum's goals stated that education must be relevant to life's needs and provide students with opportunities to apply their knowledge in society (Kemdikbud, 2012). Consequently, statistics in high school should be designed to be relevant to students' lives outside of school because they are now living in a data-driven society. Learning statistics should enable students to apply their statistical knowledge beyond the school context.

The standard competencies for high school students were formulated to achieve these goals. However, there was a change in the grade levels at which statistics should be taught in high schools according to the regulations of the Minister for Education and Culture. Before 2016, statistics was taught in Years 7 to 11, but from 2016 onward, statistics was taught only in Years 7, 8 and 12. In addition to the above changes, the standard competencies for high school students in the domain of statistics were modified slightly in 2016. Despite this minor change, the present study's review showed that the standard competencies covered students' SL knowledge and skills sufficiently in both periods. Table 3.2 lists the regulations of the Minister of Education and Culture, while Table 3.3 presents the statistics competencies students were required to achieve in 2014, and in 2016 and 2018. As shown in Table 3.3, these competencies involved both the data producer and the data consumer perspectives. Students had to learn both how to collect, process and interpret data and present it using different representations and how to describe various data presentations and communicate information from a dataset using mean, median, mode and spread.

Table 3.2

Regulations of the Minister of Education and Culture of Indonesia Regarding the Standard Competencies of High School Students in Curriculum 2013

Year	*Permendikbud* (Regulation of Minister of Education and Culture)	Grades studying statistics
2014	No. 58 of 2014 about Curriculum 2013 for junior high school	Years 7, 8 & 9
2014	No. 59 of 2014 about Curriculum 2013 for senior high school	Years 10 & 11
2016	No. 24 of 2016 about standard competencies for high school	Years 7, 8 & 12
2018	No. 37 of 2018 about standard competencies for high school	Years 7, 8 & 12

Table 3.3

Standard Competencies in Statistics for Indonesian High School Students, 2014–2018

	Standard competencies for 2014		Standard competencies for 2016 and 2018	
Year	*Knowledge*	*Skills*	*Knowledge*	*Skills*
7	Understand the techniques of presenting two data variables using tables, bar graphs, pie charts and line graphs.	Conduct experiments to gather empirical data about real problems and present them in the form of tables and graphs.	Analyse relationships between variables and understand how to present them (in tables, line charts, bar charts and pie charts).	Present and interpret data in the form of tables, line graphs, bar graphs and pie charts.
8	Understand the techniques of presenting two data variables using tables, bar graphs, pie charts and line graphs with a computer; analyse the relationships between variables.	Collect, process, interpret and present observed data in the form of tables, diagrams and graphs of two variables; identify relationships between variables.	Analyse data based on data distribution, mean, median, mode and the spread of data to draw conclusions, make decisions and make predictions.	Present and solve problems related to data distribution, mean, median, mode and the spread of data to draw conclusions, make decisions and make predictions.
9	Determine the mean, median and mode of various types of data. Choose a two-variable data presentation technique and evaluate its effectiveness; determine the relationship between variables, based on data, to draw conclusions.	Collect, process, interpret and present data in the form of tables and various types of graphs; identify relationships between variables and draw conclusions.		

Table 3.3 (continued)

	Standard competencies for 2014		Standard competencies for 2016 and 2018	
Year	*Knowledge*	*Skills*	*Knowledge*	*Skills*
10	Describe various data presentations, in the form of tables or of diagrams or plots, that are suitable for communicating information from a dataset, through the comparative analysis of various data presentations. Present data in the form of tables or certain types of diagrams or plots, according to the information to be communicated.	Present real data in the form of tables or of diagrams or plots, according to the information to be communicated.		
11	Describe and use various measures of centre and data spreads, appropriate to the characteristics of the data, through rules and formulas, and interpret and communicate them.	Present and process descriptive statistics about data into distribution tables and histograms to clarify and solve problems related to real life.		
12			Determine and analyse the measures of centre and spread presented in the form of frequency distribution tables and histograms.	Solve problems related to presenting measurement data in frequency distribution tables and histograms.

Although statistics has not been taught to students in Years 9, 10 and 11 since 2016, this has no bearing on the decision to assess the SL of Year 9 and 12 students. Recall that assessing the SL of students in Years 9 and 12 is essential as junior high school concludes in Year 9 and senior high school ends in Year 12. Moreover, the review of the curriculum revealed that Year 8 students learned statistics near the end of semester two. The expectation is that they will have retained their knowledge by the time they are assessed in Year 9. In addition, the review confirmed that students in Year 12 acquired statistics from the beginning of the first semester. This further demonstrated that when Year 12 students' SL are assessed, they only recently learned statistics after they have not in last three years.

In relation to the data consumer perspective, the assessment framework utilised in this study aligns with the standard competencies. Some of the framework's four SL skills appear as the standard competencies that students should achieve. For example, Year 7 students are required to interpret data in any type of representation, Year 8 students must learn to make decisions after solving a problem regarding the data measure of centre and Year 9 students must learn to draw conclusions from data. Students must also learn the framework's three SL components. Regarding the representations, students at all grade levels are exposed to a variety of representations, including tables, graphs and histograms. Regarding the statistical-mathematical knowledge, students learn numerous statistical ideas, such as measures of centre and spread. However, the involvement of text and context component in the statistics domain needed to be further investigated by reviewing mathematics textbooks.

3.3 Review of Indonesian Mathematics Textbooks

Although the standard competencies confirm the opportunities provided by the K13 for students to be statistically literate, a closer review of the opportunities provided by Indonesian mathematics textbooks needed to be conducted. Mathematics textbooks strongly influence mathematics teaching and learning (Büscher, 2022b; Landtblom, 2018; Ponte &

Marques, 2011). These textbooks play an important role in mathematics education (Glasnovic Gracin, 2018). Additionally, mathematics textbooks contain the pedagogical translation of government policy and show the link between the intended and implemented curriculum (Valverde et al., 2002; Weiland, 2019). Therefore, the findings of a textbook review can provide a broader and more comprehensive understanding of both curriculum requirements and classroom practices (Glasnovic Gracin, 2018). Such a review can focus on how mathematics textbooks present mathematical topics in relation to current curriculum goals (Ponte & Marques, 2011). It can also assess how textbooks provide students with opportunities to learn (Choi & Park, 2013) because textbooks are the primary resources for teachers and students (Landtblom, 2018; Reys et al., 2004).

Two important decisions to make when reviewing textbooks are which media used to review and which parts of the textbooks to review. This study reviewed textbooks utilising the SL skills and components to better understand how they might be presented to students. Among the many parts of a textbook that need to be reviewed, the present study's review focused on mathematical activities and statistical contents. Reviewing the activities in the textbooks can assess whether these activities were designed to develop students' SL skills in line with the standard competencies. Reviewing the statistical content in a textbook will also assess whether that content was designed in line with the three SL components. For review purposes, the mandatory high school (Years 7–12) mathematics textbooks published by the Indonesian Ministry of Education and Culture (e.g., As'ari et al., 2016, 2017a, 2018; Sinaga et al., 2014a, 2014b; Subchan et al., 2015) were selected. Some of these textbooks followed the 2014 regulations, while others followed the 2016 and 2018 regulations. These books were selected because they are free to download for all students across Indonesia. Although it is not compulsory for schools and students to use these textbooks, the government textbooks were considered representative of Indonesian mathematics textbooks. It is assumed that textbooks from other publishers do not differ substantially from the government ones.

In terms of the activities, the textbooks were designed to support students to be critical learners using various learning models. The indicators that students are critical learners are the 5M activities: *mengamati*, *menanya, menalar*, *mencoba* and *mengomunikasikan* (e.g., As'ari et al., 2016, 2017a, 2018). In English, these can be called the 5L activities: let's observe, let's ask, let's reason, let's find information and let's communicate. These activities relate to three learning models: discovery learning, problem-based learning and project-based learning (e.g., As'ari et al., 2016, 2017a, 2018; Sinaga et al., 2014a; Subchan et al., 2015). These 5L activities were visible in all the textbooks reviewed. However, they were clearly presented in four textbooks (for Years 7, 8, 9, and 12) that were written by the same authors. Table 3.4 summarises the activities—the 5L plus two additional activities, let's work on a project and let's summarise—and how they apply to examples found in the Year 8 mathematics textbooks.

Table 3.4

Design of Statistical Activities in Indonesian Year 8 Mathematics Textbooks

Activity		Description	Example
5L	Let's observe	This activity helps students to develop their ability to find information.	Students are provided with raw data on nine Year 8 students' weights (in kg: 47, 57, 53, 50, 45, 48, 52, 49, 55) and asked to identify the mean, mode and median.
	Let's ask	Based on the information they found, students can ask questions either on parts they do not understand or to search for further information.	Students are asked to identify the differences between the three measures of centre and to determine which one best explains the weight data. Students are also encouraged to pose questions related to mean, mode and median.

Table 3.4 (continued)

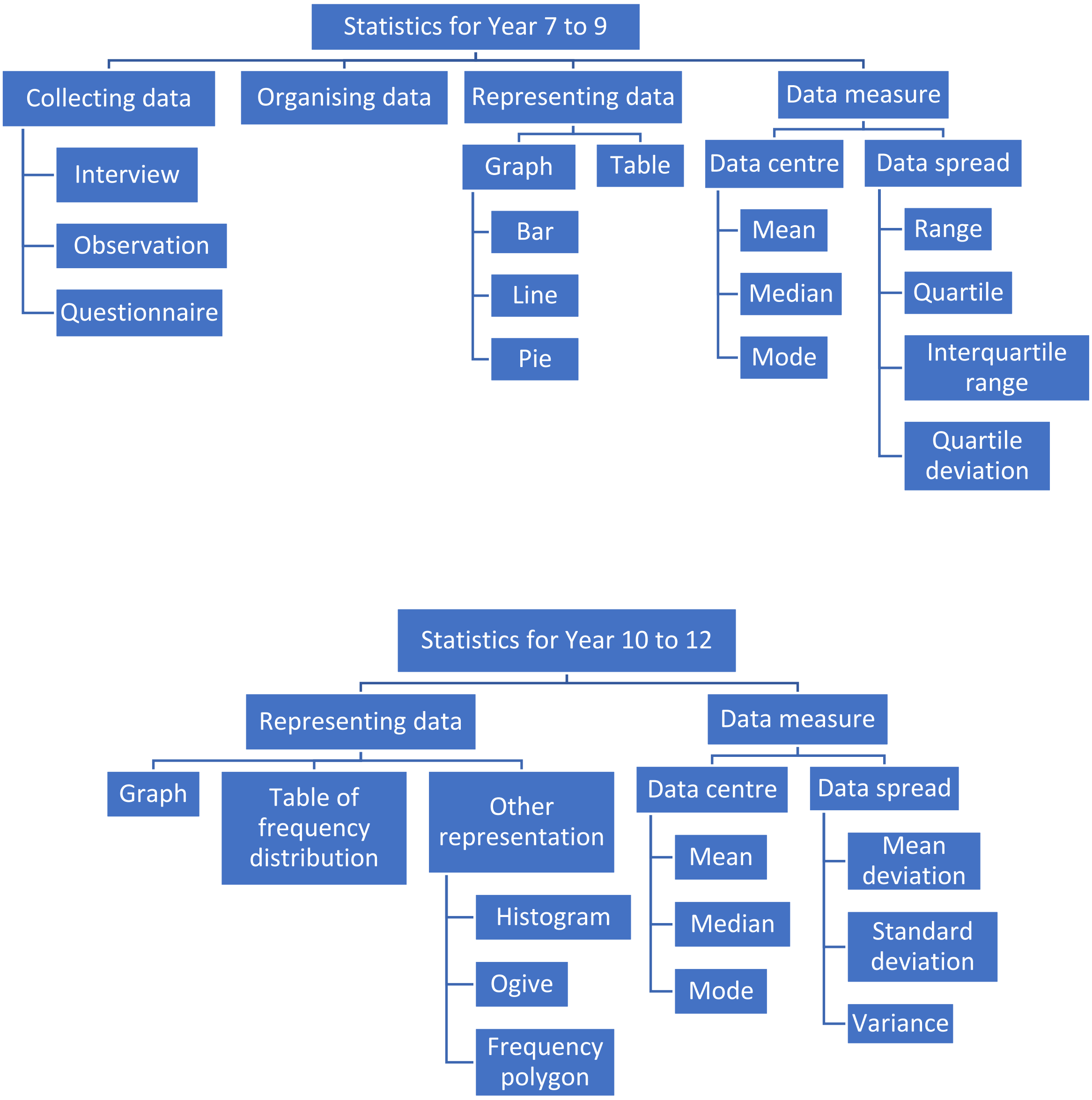
Activity		Description	Example
	Let's find information	From the questions asked, students can find more information from the book or other sources.	One case is designed in which there is a new Year 8 student joining the existing nine, and this student's weight is 51 kg. Students are asked to identify the changes, if any, applied to the three measures of centre.
	Let's reason	In this activity, students are expected to process the information they collected and provide a conclusion.	Another contextual problem is provided, and it relates to mean, mode, and median. The problem is about finding one best runner of the three runners after their time across six races have been recorded.
	Let's share	It is the time for students to share or communicate the result of their observation in either written or spoken form.	Students solve the runner problem in a group and present it in front of the class.
Additional 2L	Let's work on a project	The students are provided with a project to solve.	Students conduct a survey on how many hours their classmates watch TV in a day compared to how many hours they spend on self-study. One example question they need to solve is: which measure can be used to compare these data?
	Let's summarise	In this activity, students receive questions (from the teacher) that guide them to summarise the chapter.	Students summarise what they learned about statistics.

It can be identified from Table 3.4 that the SL assessment framework proposed in the present study (see Section 2.6) aligns with the design of the activities. In terms of SL skills, the 5L activities cover at least three skills (interpreting, communicating and decision-making). Through the let's observe activity, students learn how to interpret data under different contexts and in many forms of representation. For example, the students are asked to find the three measures of centre of a weight dataset. Further, students' communicating skill is incorporated into the let’s share activity, which encourages them to write a report and present it orally. In regard to the decision-making skill, the contextual problem in the let’s reason activity enables students to choose and provide evidence to support their choice. Although the evaluating skill might not be clearly visible, when students provide proof to support their conclusion (a claim) in the project-based activity and present it in front of the class, other students are encouraged to evaluate the claim made by the presenting group. Thus, the evaluating skill is embedded in the 5L activities.

A further review was conducted to reveal the statistical content taught in the textbooks. This review related to the framework’s three components of SL (text and context, representation and statistical-mathematical knowledge). Regardless of the regulations that the textbooks followed, all the abovementioned textbooks were reviewed. The results regarding the statistical content taught across high schools in Indonesia are presented in Figure 3.2. Overall, the statistical content covered in Years 7 to 9 is similar to that of Years 10 to 12. The major difference is in the type of data and representation. Years 7 to 9 learn about ungrouped data, while Years 10 to 12 focus more on grouped data. Further, Year 10 to 12 learn about histogram and ogive in addition to bar graphs, line graphs and table.

Figure 3.2

Statistical Contents of Indonesian High School Mathematics Textbooks



3.4 Review of Demography and Pedagogy of Indonesian Mathematics Teachers

Although the primary focus of this study is the performance of Indonesian students in statistics education, a review of the demography and pedagogy of mathematics teachers in Indonesian schools provides further context for the study. Moreover, Indonesia is contextually and culturally different from Western countries, and thus the SL might be interpreted and taught differently by the teachers. By discussing their demography, it is easier to understand the backgrounds and qualifications of mathematics teachers in Indonesia. Furthermore, it is imperative to examine their pedagogy as they engage directly with students and play a pivotal role in implementing the curriculum in schools (Noor et al., 2020b). In classrooms, these teachers are required to develop students' SL using approaches beyond formal teaching and opportunities beyond procedural calculations. Each of those is elaborated below.

School statistics are taught in Indonesian classrooms as part of mathematics (Tiro, 2017), as in other countries (Büscher, 2022a; Büscher, 2022b; Zieffler et al., 2018). In all levels of education, it is mathematics teachers who teach statistics rather than teachers with a background in statistics. These teachers graduate from a bachelor program in mathematics education (Abadi & Chairani, 2020). One of the reasons causing the absence of statistics teachers in schools might be due to the absence of statistics education majors in the bachelor program. This contrasts with mathematics which has two separate majors, namely mathematics and mathematics education. Regardless of their educational background, these mathematics teachers play a big part in teaching statistics, which is regarded as a distinct field from mathematics.

According to the Indonesian Mathematical Society (IndoMS; Abadi & Chairani, 2020), these mathematics teachers learned certain statistics courses as part of their bachelor's degree in mathematics education. Typically, those courses cover descriptive statistics, inferential statistics and probability (Yusuf et al., 2020). Nonetheless, different universities may have different names for those courses. For instance, *Statistics I* covers descriptive statistics, *Statistics II* covers inferential statistics and *Mathematics Statistics* covers statistics and probability. The first two courses would help them to analyse data for their final project or mini thesis, while the third course would help them to comprehend the uncertainty in the real world. More importantly, those three courses provide them with statistical reasoning so that they could use when teaching statistics to school students (Yusuf et al., 2020).

Given their background in three statistics courses from their bachelor's degree in mathematics education, mathematics teachers are required to be professionals. Professional and qualified teachers are needed to improve the quality of statistics instruction, students' competence and learning outcomes (Noor et al., 2020a). One of the several non-structural teacher organisations in Indonesia that supports the professional development of mathematics teachers is named Subject Teaching Working Group, or *Musyawarah Guru Mata Pelajaran* (MGMP; Abadi & Chairani, 2020). The MGMP provides a venue for high school mathematics teachers to discuss and share their experience related to problems encountered in teaching and learning mathematics, lesson plan preparation, teaching materials, approach, method, evaluation, and issue on the implementation of a new curriculum (Abadi & Chairani, 2020; Noor et al., 2020a; Noor et al., 2020b). Among the MGMP programs that aim to improve the teachers' competence and professionalism are collaborative research, training of scientific writing, curriculum analysis and discussion for material comprehension (Noor et al., 2020b).

However, despite the fact that MGMP plays a strategic role in strengthening and improving teacher competence through discussion and training (Noor et al., 2020b), the ability of mathematics teachers to think statistically is still arguably. It is debatable if these teachers show improvement after they attended many MGMP programs. The low results of their teachers' competency test, or *Uji Kompetensi Guru* (UKG; Noor et al., 2020a) and their low ability to implement K13 (Noor et al., 2020b) are indications of their lack of competence. The teachers from the unorganised MGMP tend to have such low results in the UKG, whereas teachers from well-organised MGMP tend to have high results in the UKG (Noor et al., 2020a). Although little is known for their competence in statistics, a study on pre-service mathematics teachers' competency revealed that less than 50% of them were proficient in statistical reasoning for descriptive statistics (Yusuf et al., 2020). Their poor mathematical skills and ignorance of descriptive statistics were the root causes of this deficiency (Yusuf et al., 2020). As a result, rather than employing the inductive logic of statistics, these mathematics teachers frequently teach statistics using the deductive logic of mathematics.

To support teachers’ pedagogy, teachers' textbooks that supplement students' mathematics textbooks provide extra information for teachers. These teachers’ textbooks translate the government policy into pedagogy and expected implementation. In accordance with K13, the teachers’ textbooks describe a few potential learning models, approaches and methods that can be applied in the classroom. Given that K13 aims to enhance teachers' quality as well as students' competences (Zuhdi, 2015), teachers are encouraged to apply the suggested learning models, approaches and methods (As’ari et al., 2017b; Sinaga et al., 2014c) to encourage the growth of students’ statistical thinking. There are three suggested learning models in the teachers’ textbooks: discovery learning, problem-based learning and project-based learning. Unlike traditional education, which is mostly teacher-oriented, these three learning models are student-oriented. The recommended learning approach is a scientific approach that supports students’ investigation beyond procedural calculations. Finally, the following learning methods are advised: problem solving, questioning, discussion and project.

Putting aside those different models, approaches and methods, pedagogy remains central to education (Zuhdi, 2015). The teachers must be proficient in effective learning practices in order to be competent educators as well as to understand their students (Patahuddin et al., 2018). In the case of statistics, the teachers should provide students with opportunities to explore data. The 5L activities in the mathematics textbooks provide clear instruction for teachers to implement effective learning (see again Section 3.3). Once the teachers implement the 5L activities, students' SL skills may develop as required. Otherwise, the students may not develop their ability in evaluating and communicating data-based information.

3.5 Review of UN (Indonesian National Examination), PISA and TIMSS

Although statistical content in textbooks can directly indicate what teachers teach and students learn, what is tested in the standardised tests provides another avenue for analysis, given the typical link between instruction and testing. Consequently, a review was conducted to statistics items used in the Year 9 and Year 12 UN examinations. To provide a comparison, statistical items in two other large-scale assessments in which Indonesian high school students participated were also examined. Those two international assessments were PISA and TIMSS.

The performance of students in UN can indicate their statistical knowledge and indirectly provide clues about classroom practices. It is surprising that the Year 9 and Year 12 UN results for the past few years showed that students underperformed in mathematics, including statistics (Sumaryanta et al., 2019). In 2019, 64.98% of Year 12 Science students and 44.40% of Year 9 students responded incorrectly to questions on statistics and probability, whereas in 2018, 62.53% of Year 12 Science students and 54.29% of Year 9 students answered incorrectly to questions in this domain (Pusat Penilaian Pendidikan Kemdikbud, 2023). These results suggested that Year 9 students may perform better than Year 12 students on their respective UN for statistics and probability domain. In addition, this underperformance in the UN suggests that the education system in Indonesia is not yet optimal (Sumaryanta et al., 2019).

Similar to the UN results, the majority of Indonesian students performed at a very low level in PISA and TIMSS, including in the statistics strands (e.g., Mullis et al., 2012; OECD, 2004, 2014). A possible factor influencing their low performance is the unfamiliar competencies they have to demonstrate, along with the possible impact of the format and competencies of the assessment (Anagnostopoulou et al., 2010). Therefore, the statistical items in these three large-scale assessments were reviewed to identify similarities and differences in item content and format. They were also reviewed in relation to the four SL skills and the three SL components of the present study's framework.

The identification of statistical items resulted in 82 items from the three large-scale assessments for comparison. These items included 13 items from TIMSS 2011, 12 from PISA 2012, 40 from the Year 9 UN and 17 from the Year 12 UN. Each Year 9 UN and Year 12 UN item was sourced from tests during 2008–2018 that were downloaded from unofficial sites. However, several mathematics teachers have confirmed the authenticity of these UN items. The year range for the UN tests was essential to capture the trend over the decade and to correspond with the curriculum changes and the years of TIMSS 2011 and PISA 2012 tests. The PISA and TIMSS statistical items were collected from the released items downloadable from the official websites (OECD, 2013b; TIMSS & PIRLS, 2011).

The initial focus of the review was on the proportion of statistical items in each mathematics assessment. In TIMSS, statistical items are under 'data and chance' domain and account for 20% of all items (Mullis et al., 2012). In PISA, statistical items are under the 'uncertainty and data' domain and account for 25% (OECD, 2004, 2014). Given that around 12 to 18 items were tested in TIMSS for Year 8 (Mullis & Martin, 2017), there were about three 'data and chance' items. Similarly, given that around 16 items were tested in PISA (Wijaya et al., 2014), the number of 'uncertainty and data' items was four. Moreover, these numbers covered both statistics and probability items, suggesting that the actual numbers of data-based items were less than these figures. These findings suggest that around two data-based items were included in each test, which is a relatively small number with which to assess students' understanding of data-based problems.

In comparison, the percentage of statistical items in the UN could not be obtained from official sources and had to be discerned by manually counting the items in each test. In the UN tests from 2008 to 2018, data-based items (excluding probability items) accounted for a smaller but increasing proportion compared to items from the other strands of mathematics. For Year 12, the proportion increased from 2.5% in the 2008 test to 5% in the 2016 and 2017 tests and 7.5% in the 2018 test. This is because the number of statistical items increased from one to three items in the past decade in the Year 12 UN. For Year 9, the proportion of statistical items in the test was larger than that for Year 12, being either 7.5% or 10%. The proportion remained stable at 10%, or four items, in the 2016 to 2018 tests. Despite the increase, these proportions remained relatively small. Therefore, assessing students' understanding of statistics based on the UN is also not sufficient.

In addition, the statistical content and item formats used in the UN were very different from those in PISA and TIMSS. In general, the competencies tested in the Year 9 UN related to data representations and measures of centre and spread. The Year 12 UN also tested such competencies but expanded them to involve quartiles, histograms, grouped data and finding the mode, median and mean of grouped data. This finding suggests that, over the decade, similar items and content were repetitively used regardless of the curriculum changes. Further, Year 9 and Year 12 UN both differed from TIMSS and PISA, which required students to engage in higher-order thinking. This finding supports that of Winarti and Patahuddin (2017), who showed that the different aims of these large-scale assessments resulted in different items being tested. Table 3.5 presents the statistical content tested in the UN. Following on, Table 3.6 shows the item format distributions of these three large-scale assessments. All the UN items were tested in multiple choice formats, in contrast to the more varied formats of TIMSS and PISA. In addition, TIMSS and PISA also have long-answer problems and always use data in a representation format, while not all problems in the UN use data in this way.

Table 3.5

Statistical Contents Tested in UN

	Year 9	Year 12
Knowledge	Students can understand knowledge about the following areas: • presenting and describing data in the form of a table, bar graph, line graph or pie chart • measuring the centre of the data	Students can understand the following basic concepts: • presenting data in the form of a table, diagram or graph • measuring the centre of the data and its location and spread
Application	Students can apply knowledge about the following areas: • presenting and describing data in the form of a table, bar graph, line graph or pie chart • measuring the centre of the data	Students can apply statistics concepts in the following contextual problems: • presenting data in the form of a table, diagram or graph • measuring the centre of the data and its location and spread
Reasoning	Students can use reasoning related to the following areas: • presenting data in the form of a table, bar graph, line graph or pie chart • measuring the centre of the data	Students have reasoning ability in the following areas: • presenting data in the form of a table, diagram or graph • measuring the centre of the data and its location and spread

Table 3.6

Format and Features of Statistical Items in Three Large-Scale Assessments

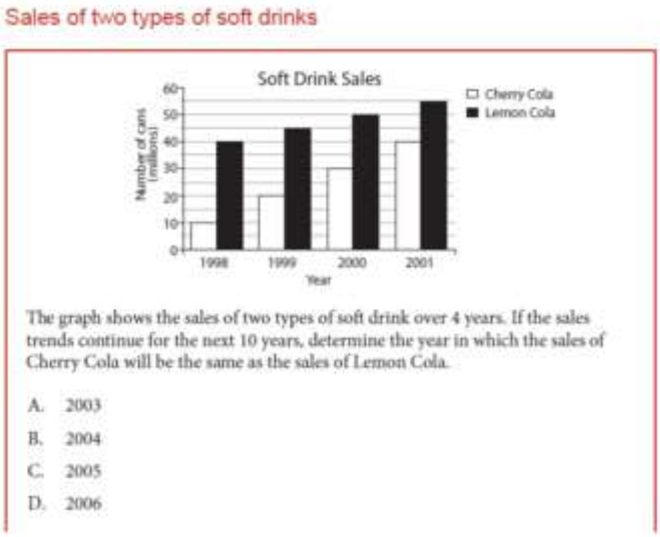
	TIMSS	PISA	Year 9 UN	Year 12 UN
Format				
Multiple choice	3	6	40	17
Short answer	6	–		
Long answer	2	3		
Yes/no	–	3		
Construct a representation	2	–		
Feature				
With representation	13	12	25	14
Without representation			15	3

Note. TIMSS is Trends in International Mathematics and Science Study; PISA is Programme for International Student Assessment; UN is *Ujian Nasional* [Indonesian National Examination].

A further review was conducted on these 82 items to reveal the skills and components that were included in the tests. This classification aimed to better understand to what extent the four SL skills were assessed through the large-scale assessments. The review was conducted by reclassifying the statistical items from their original constructs into the four SL skills: interpreting, communicating, evaluating and decision-making. For example, Figure 3.3 shows the Soft Drink problem from TIMSS and the Charts problem from PISA. Based on their respective frameworks, the Soft Drink problem was categorised into the ‘reasoning’ domain under the TIMSS framework, while the Charts problem was classified under the ‘employing’ process in the PISA framework. Since these two items are similar, a cross-check was undertaken before they were reclassified. The Soft Drink problem was considered a ‘reasoning’ problem because students needed to solve it using logical thinking based on a pattern called a ‘trend’. The Charts problem also provided a trend for students to use as a clue to estimate the heights of bars outside the given times. Therefore, ‘trend’ was used as the keyword.

Figure 3.3

Example of Items Recoded as Interpreting Items



Sales of two types of soft drinks

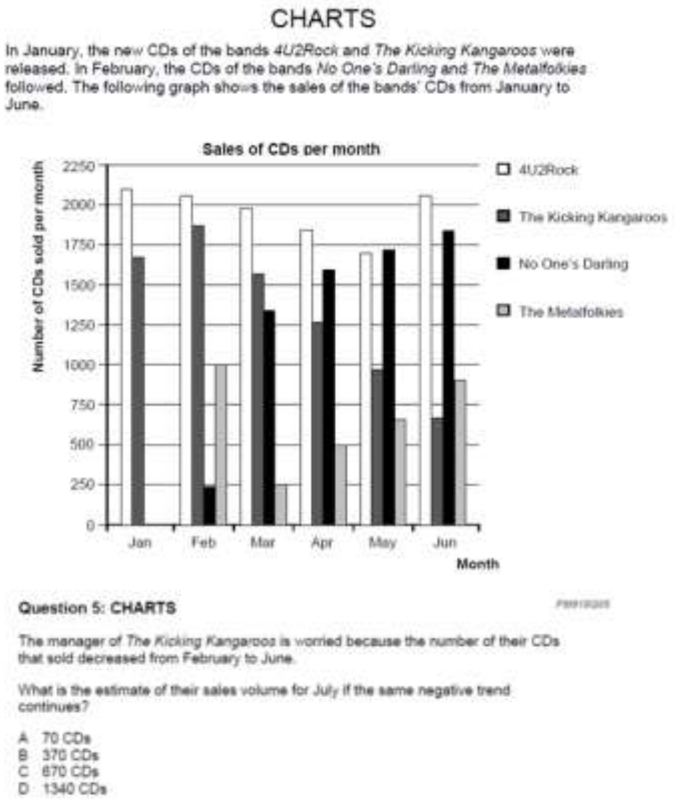
The graph shows the sales of two types of soft drink over 4 years. If the sales trends continue for the next 10 years, determine the year in which the sales of Cherry Cola will be the same as the sales of Lemon Cola.

A. 2003

B. 2004

C. 2005

D. 2006



CHARTS

In January, the new CDs of the bands *4U2Rock* and *The Kicking Kangaroos* were released. In February, the CDs of the bands *No One's Darling* and *The Metalfolkies* followed. The following graph shows the sales of the bands' CDs from January to June.

Question 5: CHARTS

The manager of *The Kicking Kangaroos* is worried because the number of their CDs that sold decreased from February to June.

What is the estimate of their sales volume for July if the same negative trend continues?

A 70 CDs
B 370 CDs
C 670 CDs
D 1340 CDs

Note. Soft Drink is a TIMSS released item (TIMSS & PIRLS, 2011); Charts is a PISA released item (OECD, 2013b)

By referring to the definition of the four skills in Table 2.6 (see Chapter 2) and using ‘trend’ as the keyword, the review concluded that these items assessed an interpreting skill in which students explained the implicit meaning of a trend and used logical reasoning to assess how the trend affected the bars. Moreover, these items were considered to assess interpreting skills for a further reason: because they do not ask students to share their opinions (communicating), challenge a statistical claim (evaluating) or make a choice (decision-making).

This reclassifying process was applied to all 82 statistical items and was consulted with one expert. The results are presented in Table 3.7. The findings from reclassifying typical items in the three large-scale assessments showed that the UN assessed only one of the four skills, whereas PISA and TIMSS assessed more skills. Moreover, all the statistical items in the Year 9 and Year 12 UN only measured basic interpreting skills that required students to know about basic descriptive statistics in familiar contexts, about simple representations and about suitable calculation procedures. However, one statistical item in the Year 12 UN 2014 was categorised as a PISA-like problem. This finding might explain the underperformance of Indonesian students in PISA, since it was found that approximately 73% of Indonesian students could only read data from a simple graph or table, only 23% could solve a problem by interpreting data from any kind of representation, and only 2% could successfully solve statistical problems with more complex contexts and skills, such as evaluation skills (OECD, 2014).

Table 3.7

Distribution of Statistical Items in Large-Scale Assessments Across the Four Skills

	PISA	TIMSS	Year 9 UN	Year 12 UN
Interpreting (I)	8	11	40	17
Communicating (C)	1	2		
Evaluating (E)	2			
Decision-making (D)	1			

Note. TIMSS is Trends in International Mathematics and Science Study; PISA is Program for International Student Assessment; UN is *Ujian Nasional* [Indonesian National Examination].

In summary, the findings of this review concluded that almost all the items in the UN assessed only basic interpreting skills, whereas TIMSS assessed interpreting and communicating skills, and PISA assessed all four skills. Moreover, the UN, which focused on basic interpreting skills, provided questions only in a multiple-choice format and used repetitive statistical content over the years for both Year 9 and Year 12. The competencies tested in the Year 9 UN were basic statistical knowledge, including reading data from representations and finding the mode, median and mean from data in representations. The Year 12 UN also assessed such competencies but expanded the assessment to include quartiles and histograms. These findings suggest that the UN was unsuitable for monitoring the improvement of Indonesian students' SL. However, TIMSS and PISA can likely be used as a reference since both tests offer some statistical items covering different SL skills and item formats, even though there were only a limited number of data-based items in each test.

3.6 Chapter Summary

This chapter has provided a sufficient understanding of the education system in Indonesia, specifically the Indonesian high schools. Indonesian high schools are divided into two: junior high school, which concludes in Year 9, and senior high school, which concludes in Year 12. As with other categories of education, high schools in Indonesia were administered under the auspices of two ministries (MoEC-RT and MoRA), which can be further referred to as school type. In addition, these schools have either private or public status which can be further referred to as school status. Both public and private schools are major providers of education services in Indonesia. In light of the foregoing, it is reasonable that the present cross-sectional study was also designed to investigate students' SL from various cohorts, including school types and statuses, in addition to grade levels.

This chapter has also provided a sufficient understanding of the opportunities for conducting an assessment study on SL with the Indonesian high school students. These opportunities were reflected through the *Kurikulum* 2013 (K13), which aims to help students acquire competencies aligned with the four SL skills. This curriculum was implemented to improve the quality of Indonesian education considering the underperformance of Indonesian students on international tests (such as PISA and TIMSS). In addition, mathematics textbooks provide opportunities for students to become statistically literate by including activities aligned with four SL skills, and statistical contents in line with the three SL components. Nevertheless, this curriculum was officially implemented in 2013, and there was insufficient evidence to monitor its impact on students' SL levels. The results of Year 9 and Year 12 UN cannot be used to monitor the students' SL because similar items and contents were repeatedly used over a decade, despite curriculum changes, and because UN items assess only basic interpreting skill. Moreover, the most recent PISA results, which were used to inform the SL of 15-year-old students at the time this study was conducted, were published in 2014, and there were no comparable empirical results on the SL of Year 12 students to PISA and TIMSS. Consequently, conducting this study using the proposed SL framework could provide a current perspective of the Indonesian high school students' SL.

Given the information provided in this chapter, which shows opportunities to assess Indonesian high school students using the proposed assessment framework, the following two chapters focus on the methods. Chapter 4 specifically concentrates on the validation of the framework through instrument development and piloting, while Chapter 5 emphasises more on the research methodology. Although instrument development and piloting are parts of the methodology, it was argued that they should precede data collection and presented in different chapters. This separation was intended to prevent a mix and mismatch between the methods of validating the assessment framework involving instrument development and the methods of data collection and analysis. Moreover, there were some technical terminologies related to instruments that cannot be directly mentioned in the methodology before being clearly explained in the instrument development and piloting chapter. Otherwise, the instrumentation and data analysis in Chapter 5 cannot be explained in detail using validated items.

Chapter 4: Instrument Development, Piloting and Refinement

The Introduction and Literature Review chapters have provided information about the scope of the study and the SL assessment framework, while the Study Context chapter has highlighted the potential for conducting this study with Indonesian high school students. The present chapter describes the stages of developing, piloting and revising the instruments this study used to assess Indonesian high school students' SL, which also contribute to the validation of the assessment framework. The validated framework and instruments are then used to measure students' SL levels, challenges and understanding. Section 4.1 describes the theories underpinning the instruments' development, piloting and refinement. Section 4.2 describes the initial stages of developing the instruments (the assessment items, the cognitive interview protocol and the scoring guide descriptors). Section 4.3 describes how the initial version of the assessment items and interview protocol were then piloted for refinement. This section describes all the piloting stages used to determine whether the students understood the items well, whether the questions prompted students to recall the cognitive processes they used when solving the problems and whether the descriptors for each of the six hierarchical levels reflected the students' various responses. Section 4.4 presents the test items. Finally, Section 4.5 concludes this chapter with a summary.

4.1 Theoretical Underpinnings

A careful and systematic process of developing an assessment starts with a meticulous description of the construct to be assessed (Ralston et al., 2018). Although living in a data-driven society clearly requires citizens (including students) to possess a variety of essential data consumption skills, SL is a broad concept, and researchers have widespread disagreements about its definition and construct. Therefore, as described in Chapter 2, the definitions of SL provided by previous researchers were reviewed (e.g., Budgett & Renelle, 2023; Budgett & Rose, 2017; Büscher, 2022a; Gal, 2002; Mullis et al., 2012; OECD, 2014; Wallman, 1993; Watson & Callingham, 2003) along with the various skills researchers have identified as necessary for students to be data consumers. Chapter 2 also describes how that review resulted in the SL assessment framework for the current study. The framework includes constructs and sub-constructs. The constructs comprise four response skills (interpreting, communicating, evaluating and decision-making); each skill comprises sub-constructs involving three knowledge components (text and context, representation and statistical-mathematical knowledge). These constructs and sub-constructs were used as the basis for this study's instrument development.

It is essential to use quality instruments to measure students' learning (Sabbag et al., 2018) and this chapter concentrates on describing the development of such an instrument prior to data collection. Based on the SL framework, the stages of developing the instruments—assessment items, cognitive interview protocols and component-based item descriptors used to characterise students' SL—were defined. For item development, this study adapted DeWalt et al.'s (2007) and Van den Heuvel-Panhuizen's (1996) methods to produce initial items appropriate to the SL framework in terms of assessed skills and components. DeWalt et al. (2007) applied a method called qualitative item review (QIR) that included item identification, item classification and selection, item revision, focus group exploration of domain coverage, cognitive interviews and, finally, revision. In comparison, Van den Heuvel-Panhuizen (1996) applied a method consisting of three major stages: generation, selection and adjustment. These three stages align with the first three stages of QIR: item identification corresponds with generation, item classification and selection correspond with selection, and item revision can be considered similar to adjustment. The terms in Van den Heuvel-Panhuizen's (1996) method, however, were used in this study.

To complement item development, level descriptors and a cognitive interview protocol were developed to facilitate an SL test and an interview, respectively. For the level descriptor development, the process used the component-based descriptors from Chapter 2. This Chapter 4 elaborates on those descriptors in relation to each item and provides the initial component-based item descriptor. Meanwhile, the initial cognitive interview protocol was developed using the four processes of Shafer and Lohse (2005) and Willis (1999, 2005): the students' comprehension of the problem, their processes to retrieve relevant information, their decision processes and their response processes. By exploring the above processes, the cognitive interview aimed to ensure that the questions measure the intended construct, are correctly interpreted as well as to investigate students' general understanding and misconceptions when solving SL items (Reinhart et al., 2022). Additionally, the interview helped the researcher examine whether the students performed one of the four skills — interpreting, communicating, evaluating and decision-making—using the three knowledge components, when they were asked to do so.

After the initial versions of the instruments were finalised, three-stage piloting was conducted to ensure the instruments' applicability. Instrument piloting has broader purposes, as explained by Tiruneh et al., 2017. First, it aims to provide evidence of whether the assessment items are clear to the respondents. Second, it aims to examine whether the expected responses can be obtained from students under the test conditions. Third, it aims to test whether the scoring guide can be used. Finally, it aims to test whether the interview protocol can prompt students to recall the thought processes they used when solving more complex problems. Considering these aims, the three-stage piloting in this study was designed to ensure the applicability of the three instruments together instead of separately. The three pilot stages were Pilot Interview I, the Pilot Test and Pilot Interview II. Pilot Interview I, remarkably efficient at creating rich qualitative data when examining individual items with a small sample of respondents (DeWalt et al., 2007), was used to check the applicability of the assessment item and the interview protocol. Then, the Pilot Test was used to check the applicability of the assessment items and the descriptors, and Pilot Interview II was used to check the applicability of all three instruments.

After that, based on the findings from the three pilot stages, the final refinements to the assessment instruments (items, descriptors and interview protocol) were made. These final refinements aimed to minimise measurement error. Ziegler and Garfield (2018) mentioned four measurement errors, summarised from Weathington et al. (2010), including instrument error, participant variability, researcher variability and environmental variability. Minimising instrument error means minimising wording and organisational issues; this applied to all three instruments after piloting. Minimising participant variability relates to reducing participants' fatigue and misunderstanding of items; this applied to the items and the interview protocol. Minimising researcher variability relates to reducing recording errors, and this applied to the component-based item descriptors. Lastly, minimising environmental variability means reducing distractions and differences in testing locations; this was ensured throughout piloting and was considered during the actual data collection.

Finally, the final version of the assessment instruments for assessing students' SL—in line with the proposed assessment framework—was obtained. The results of the pilots confirmed the applicability of the assessment framework. More importantly, the pilot results suggested that the refined instruments could successfully be used to measure students' SL levels and to investigate the challenges students encounter and their understandings. Each stage of the instrument development is described in the subsequent sections.

4.2 Initial Stage of Instrument Development

This section focuses on the initial instrument development process along with the results of that process. The initial instrument development was conducted from January to

July 2019. Section 4.2.1 presents the process of developing assessment items, as well as the resulting items; Section 4.2.2 presents the interview protocol development; and Section 4.2.3 presents the development of the component-based item descriptors.

4.2.1 Results of Item Development

This section presents the results from the initial stage of item development. There were three steps in this initial stage: generation, selection and adjustment. Table 4.1 summarises these stages and the obtained results for each stage. Explanations of each stage are given in the following sections.

Table 4.1

Initial Stages of Assessment Item Development

Stage	Description	Output
Generation	Analyse the Indonesian mathematics curriculum, mathematics textbooks and the Indonesian national examination in mathematics	Curriculum goals; standard competencies for statistics; typical statistical tasks; the statistical content to be included in the SL test
Selection	Analyse some data-based items from different sources and select the potential items	14 selected items
Adjustment	Adapt the selected items	The adapted 14 items

4.2.1.1 Generation

The results of the generation process, which started with curriculum review, were mostly presented in Chapter 3; consequently, the present section exclusively focuses on which statistics content needs to be included in the SL test. The results of the comprehensive review of the curriculum, mathematics textbooks and UN (Indonesian National Examination) in Chapter 3 guided the selection of statistics content in this section. Particularly, the selected statistics content needed to align with the curriculum goals, the essential statistics competencies expected to be achieved by students and the statistical content already being taught and tested; all of which were described in Chapter 3. More importantly, the content needed to be appropriate for Indonesian Year 9 students (approximately 15 years old) and Indonesian Year 12 students (approximately 18 years old). Such consideration of the appropriate content for both grade levels is necessary because this study employed a cross-sectional design (see Chapter 5).

After examining and comparing the statistical content of the textbooks and the UN (see Table 4.2), two areas—data measures of centre and spread, and graphical and tabular representations of ungrouped data—were eventually selected. Year 9 and Year 12 students would both have studied this content. Students in Years 7–9 and Year 10 were learning about ungrouped data, and consequently, in the present study's instruments, data measures of centre and spread were set in the context of ungrouped data. In addition to simple graphs and tables, these students learn about complex graphs; thus, double-bar graphs and graphs with discontinued axes were included. Histograms and ogives were not used because they were being taught only to students in Years 11 and 12.

Table 4.2

Statistical Content in the Textbooks and UN

	Statistical content in textbooks	Statistical content tested in the UN
Years 7–9	Students learn how to collect, process, interpret and present observed data using tables, bar graphs, line graphs and pie charts of two variables and identify relationships between variables. They also learn to choose the most effective data presentation techniques, determine relationships between variables based on data and draw conclusions. Students learn how to determine the mean, median and mode of various types of ungrouped data. Further, they learn to analyse data based on data distribution, mean, median, mode and the spread of data (range, quartile, interquartile range and quartile deviation) to draw conclusions and make decisions and predictions.	Students are assessed on their ability to understand, apply and use reasoning based on: • the ungrouped data presented as a table, bar graph, line graph or pie chart. • the data measures of centre.

Table 4.2 (continued)

Years 10–12	Students learn how to describe various data presentations suitable for communicating information from ungrouped datasets (tables, bar graphs, line graphs and pie charts) through comparative analysis of various data presentations. Additionally, they learn about histograms, ogives and frequency polygons for grouped data. Students learn how to determine and analyse descriptive statistics of ungrouped and grouped data, using distribution tables and histograms to clarify and solve problems related to real life. The data spread measures include mean deviation, standard deviation and variance.	Students are assessed on their ability to understand, apply and use reasoning based on: • the ungrouped and grouped data presented as a table, bar graph, line graph or histogram. • the data measures of centre, location and spread.
Similarities between Years 7–9 and Years 10–12	Students learn how to present and interpret tables, bar graphs and line graphs of ungrouped data. Students also learn how to determine and analyse the measures of centre and spread for ungrouped data.	Students are assessed on their ability to interpret data in tables, bar graphs and line graphs. Students are assessed on their ability to determine and analyse the measures of centre, location and spread of ungrouped data.

Note. UN is *Ujian Nasional* (Indonesian National Examination)

4.2.1.2 Selection

The second stage, selection, is the process of selecting items by considering the item's content and format as well as the item's features and its appropriateness to the framework. Some potential items from various sources that fall under the statistics content appropriate for Years 9 and 12 students were selected. These items were sourced from existing instruments (Sabbag et al., 2018), such as international standardised assessments (i.e., PISA and TIMSS), SL assessment studies, Indonesian mathematics textbooks and government reports. The decision to utilise items from PISA, TIMSS and journals was made because these items have high validity. These sources also have large collections of statistical items in constructed response form to encourage students to write long answers. Long answers from students were crucial for this study's analysis, particularly when characterising the level of students' engagement with the three SL components. In addition to the items discovered, real data from Indonesian government reports were used to develop new items.

To ensure items' appropriateness to the framework, each of the selected items must assess one of the four skills (interpreting, communicating, evaluating and decision-making). Moreover, the items should have features related to the three SL components (text and context, representation and statistical-mathematical knowledge). This consideration of the three SL components as item features functioned as a reference when analysing the students' responses. Table 4.3 presents the 14 items, of which 11 originated from various sources and three were developed from government reports. Further, for each skill, Table 4.4 gives examples of these items and provides details about these examples, including their features and their origins (for all remaining items, see Appendix A).

Table 4.3

Initial Items Selected from Various Sources

Skill	Selected item	Source
Interpreting	Car Production Graph 1	TIMSS released items
	Car Production Graph 2	TIMSS released items
	Charts 1	PISA released items
	School Students 1	Data from government reports
Communicating	Domestic Waste	Data from government reports
	Charts 2	PISA released items
Evaluating	Faulty Players	PISA released items
	Test Scores	PISA released items
	Robberies	PISA released items
	Sport Shoes 1	Mathematics textbooks
	School Students 2	Data from government reports
Decision-making	The 100-Metre Race	Journal and mathematics textbooks
	Which Car?	PISA released items
	Sport Shoes 2	Mathematics textbooks

Note. TIMSS is Trends in International Mathematics and Science Study; PISA is Programme for International Student Assessment.

Table 4.4

Examples of Assessment Items Selected Based on the Four Skills

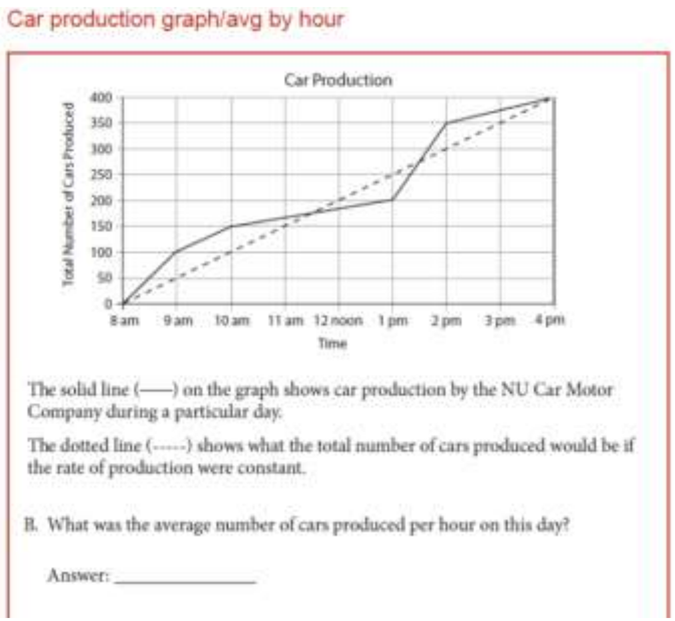
The selected item and assessed skill	Item features	Item source and format
Interpreting item (I)		
Car Production Graph 1  Car production graph/avg by hour The solid line (——) on the graph shows car production by the NU Car Motor Company during a particular day. The dotted line (-----) shows what the total number of cars produced would be if the rate of production were constant. B. What was the average number of cars produced per hour on this day? Answer: ______________	*Context:* Car production over eight hours *Representation:* A line graph presenting the increase in car production *Statistical-mathematical concepts*: Average and number operations	This item was found among the TIMSS released mathematics items. This item was considered appropriate for assessing students' ability to interpret the average from data presented in a line graph. Regarding item format, although this item was originally in a short-answer format, it could be modified into a long-answer format by asking students to explain their answers. Thus, students would be expected to successfully interpret the data displayed in a line graph based on the context (car production) to find the average number of cars produced per hour.

Table 4.4 (continued)

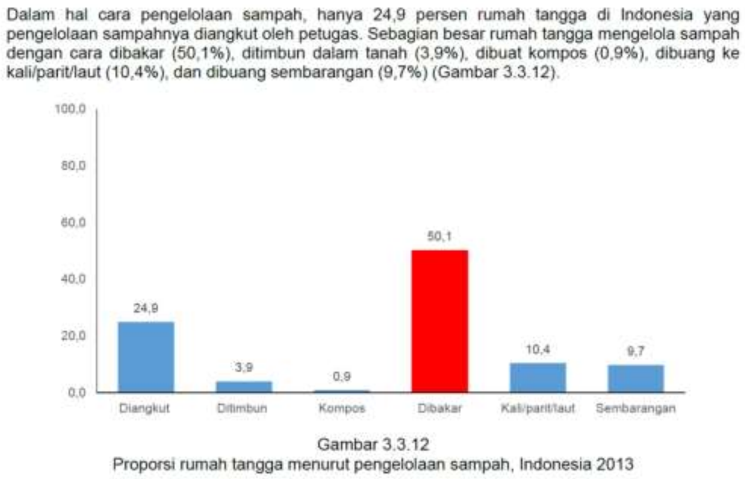
The selected item and assessed skill	Item features	Item source and format
Communicating item (C)		
Domestic Waste Translation: In terms of domestic waste management methods, only 14.9 percent of households in Indonesia have their waste transported by officers. Most households manage domestic waste by burning it (50.1%), dumping it in the ground (3.9%), composting it (0.9%), dumping it in a river, ditch or sea (10.4%) or littering (9.7%).  Dalam hal cara pengelolaan sampah, hanya 24,9 persen rumah tangga di Indonesia yang pengelolaan sampahnya diangkut oleh petugas. Sebagian besar rumah tangga mengelola sampah dengan cara dibakar (50,1%), ditimbun dalam tanah (3,9%), dibuat kompos (0,9%), dibuang ke kali/parit/laut (10,4%), dan dibuang sembarangan (9,7%) (Gambar 3.3.12). Gambar 3.3.12 Proporsi rumah tangga menurut pengelolaan sampah, Indonesia 2013	*Context*: Domestic waste management in Indonesia *Representation*: A bar graph showing six different waste management processes with contrasting percentages *Statistical-mathematical concepts*: Maximum–minimum, percentage, comparison, grouping and ordering	This report was obtained from the national report of the Indonesian Ministry of Health in 2013 (Badan Litbangkes RI, 2013). This report was considered appropriate for assessing students' ability to communicate the most relevant information from data. The report presented six actions taken by Indonesian households on their domestic waste. Because of Indonesia's littering problem, this domestic waste management is an important context. Regarding item format, a long-answer question could be constructed by asking students to summarise the Indonesian people's awareness of domestic waste. Through this, students would be expected to demonstrate critical thinking when summarising the most important features, by grouping and comparing using percentages.

Table 4.4 (continued)

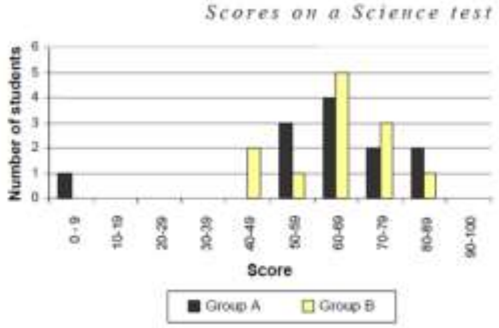
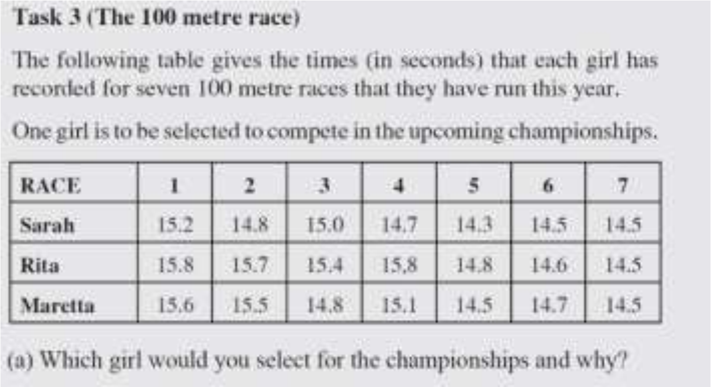
The selected item and assessed skill	Item features	Item source and format
Evaluating item (E)		
Test Scores **Question 1: TEST SCORES** M513Q01 - 0 1 9 The diagram below shows the results on a Science test for two groups, labelled as Group A and Group B. The mean score for Group A is 62.0 and the mean for Group B is 64.5. Students pass this test when their score is 50 or above.  Looking at the diagram, the teacher claims that Group B did better than Group A in this test. The students in Group A don't agree with their teacher. They try to convince the teacher that Group B may not necessarily have done better. Give one mathematical argument, using the graph, that the students in Group A could use.	*Context*: Science test scores for two groups of students *Representation*: A double-bar graph showing the distribution of students' scores in two groups *Statistical-mathematical concepts*: The minimum passing score and the effect of the outlier on the mean	This item was found in PISA released mathematics items and deemed appropriate to assess students' evaluating skills. In this item, there is a claim made by a science teacher. The question asks the students to challenge the claim. Regarding item format, this is a long-answer item asking students to utilise data in the bar graph as proof for challenging the given claim. Thus, students would be expected to demonstrate evaluating skills when challenging the claim.

Table 4.4 (continued)

<table>
<tr><th>The selected item and assessed skill</th><th>Item features</th><th>Item source and format</th></tr>
<tr><td>Decision-making item (D)</td><td></td><td></td></tr>
<tr><td><i>The 100-Metre Race</i>

<b>Task 3 (The 100 metre race)</b>
The following table gives the times (in seconds) that each girl has recorded for seven 100 metre races that they have run this year.
One girl is to be selected to compete in the upcoming championships.
<table>
<tr><th>RACE</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr>
<tr><td><b>Sarah</b></td><td>15.2</td><td>14.8</td><td>15.0</td><td>14.7</td><td>14.3</td><td>14.5</td><td>14.5</td></tr>
<tr><td><b>Rita</b></td><td>15.8</td><td>15.7</td><td>15.4</td><td>15.8</td><td>14.8</td><td>14.6</td><td>14.5</td></tr>
<tr><td><b>Maretta</b></td><td>15.6</td><td>15.5</td><td>14.8</td><td>15.1</td><td>14.5</td><td>14.7</td><td>14.5</td></tr>
</table>
(a) Which girl would you select for the championships and why?</td>
<td><i>Context</i>: A 100-metre running competition
<i>Representation</i>: A table showing the recorded times of three runners for seven races
<i>Statistical-mathematical concepts</i>: Average, maximum–minimum value and trend</td>
<td>This item was found in a published paper by Sharma et al. (2012) and is considered appropriate for assessing students’ decision-making skills. A similar item was also found in the Indonesian mathematics textbook (As’ari et al., 2017a, p. 252). Given the recorded times for seven races, the students were asked to select the best runner out of three.
Regarding item format, this is a long-answer item asking students to choose one runner and provide a justification to support their choice. Thus, students would be expected to demonstrate critical thinking when choosing the best runner and provide reasonable proof from the table.</td></tr>
</table>

4.2.1.3 Adjustment

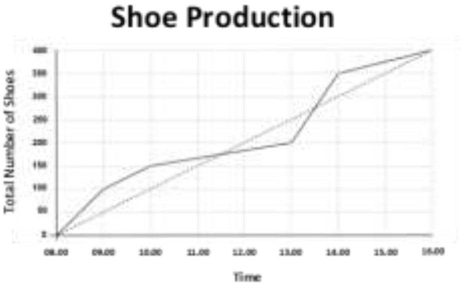
In the adjustment step, the selected items were evaluated to determine how they could be changed, and each item’s appropriateness for testing each skill was reviewed. The adaptations included changing or modifying the context, the graphical or tabular representation, the questions or the language (Tiruneh et al., 2017). Item adaptation was intended to better align the items with the research aim and participants’ context. For example, the context in some items was changed due to the participants’ unfamiliarity with the original context and the data in some items were modified to capture all possible responses from students at each of the hierarchical levels. Some questions were redesigned to require long-answer responses, aiming to prevent students from providing a short or blank answer. Table 4.5 exemplifies one adapted item for each skill, and Table 4.6 summarises the adaptations.

Table 4.5

Example of Adapted Assessment Items

<table>
<tr><th>Adapted item (translated into English)
and assessed skill</th><th>Item description</th></tr>
<tr><td colspan="2">Interpreting item (I)</td></tr>
<tr><td>

Shoe production



The solid line (—) on the graph shows the number of shoes produced by a home industry during a particular day.

The dotted line (- -) shows what the total number of shoes produced would be if the rate of production were constant.

What was the mean number of shoes produced per hour? Explain how you got it.

</td><td>

Context: The original context was *car production*; this was changed to *shoe production* to make it more familiar to the students, particularly the Year 9 students. As a result, the production place was also changed from a factory to a home industry.

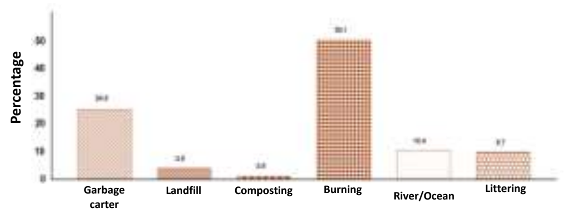
Graph: Due to the altered context, some of the graph's elements were also changed, including the title and the labels of the two axes. The y-axis now represents the number of shoes produced, while the x-axis scale was adjusted to reflect Indonesian time (i.e., 24-hour format).

Question: A slight modification was made to the question. The original version asked about the *average*, but it was changed to *mean* in this version. This change aimed to omit misconceptions as there is no exact translation of *average* in Indonesian that high school students would know.

</td></tr>
<tr><td colspan="2">Communicating item (C)</td></tr>
<tr><td>

Domestic waste

Domestic waste management in Indonesia, 2013



To make your friend informed, summarise the important information from the graph about the Indonesian people's awareness of domestic waste management!

</td><td>

Context: The context, how Indonesians manage their domestic waste, was retained, and a title was added.

Graph: The graph's elements were left unchanged, except for the bar colours.

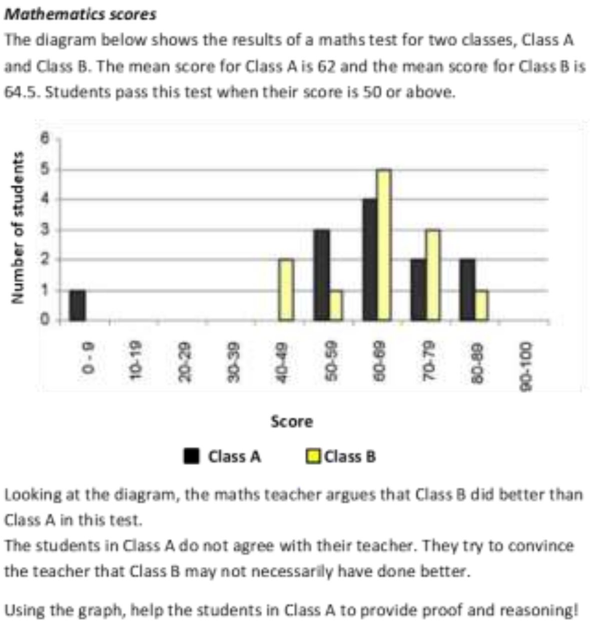
Question: As this item was intended to assess students' communication skills, a question was developed. The question asked students to summarise the important information as if they were explaining it to their friends and making them understand it.

</td></tr>
</table>

Table 4.5 (continued)

Adapted item (translated into English) and assessed skill	Item description

Evaluating item (E)



Mathematics scores

The diagram below shows the results of a maths test for two classes, Class A and Class B. The mean score for Class A is 62 and the mean score for Class B is 64.5. Students pass this test when their score is 50 or above.

Looking at the diagram, the maths teacher argues that Class B did better than Class A in this test.
The students in Class A do not agree with their teacher. They try to convince the teacher that Class B may not necessarily have done better.

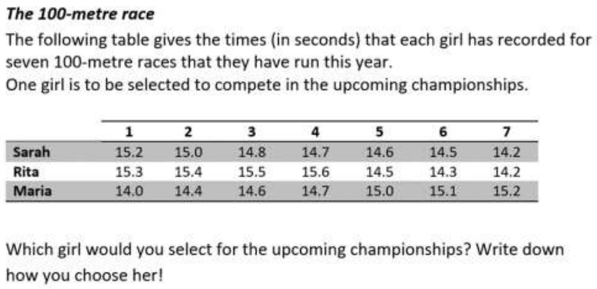
Using the graph, help the students in Class A to provide proof and reasoning!

Context: There was a slight change to the context, from science to mathematics and from two groups of students to two classes.

Graph: The graph's elements were retained, except that the legend was changed because the context was different.

Question: The question was retained because it already reflected and assessed evaluating skills.

Decision-making item (D)



The 100-metre race

The following table gives the times (in seconds) that each girl has recorded for seven 100-metre races that they have run this year.
One girl is to be selected to compete in the upcoming championships.

	1	2	3	4	5	6	7
Sarah	15.2	15.0	14.8	14.7	14.6	14.5	14.2
Rita	15.3	15.4	15.5	15.6	14.5	14.3	14.2
Maria	14.0	14.4	14.6	14.7	15.0	15.1	15.2

Which girl would you select for the upcoming championships? Write down how you choose her!

Context: The context was retained.

Table: The data in the table were modified to challenge the students more. According to a previous study using this item, most students used the mean to determine the best runner. As a result, the data were modified so that there were two runners with the same mean. However, these runners have different trends.

Question: The question was retained.

Table 4.6

Assessment Item Adaptation

Skill	Original item	Adapted item	Features changed
Interpreting	Car Production Graph 1	The Production Mean	Context, graph and question
	Car Production Graph 2	The Most Production	Context, graph and question
	Charts 1	In Which Month?	Context, graph and question
	School Students 1	How Many Students?	Graph and question
Communicating	Domestic Waste	Domestic Waste	Graph and question
	Charts 2	YouTube Viewers	Context, graph and question
Evaluating	Faulty Players	Faulty Electronics	Context, graph and question
	Test Scores	Mathematics Scores	Context and graph
	Robberies	The Employees	Context, graph and question
	Sport Shoes 1	Average Size	Text and question
	School Students 2	Dramatic Decline	Context, graph and question
Decision-making	The 100-Metre Race	The 100-Metre Race	Table
	Which Car?	Which Motorcycle?	Context, table and question
	Sport Shoes 2	More Stock	Text and question

After the adaptation process was completed, three mathematics education researchers—as expert validators—checked the modified items and the skill assessed by each item. This checking was intended to observe how experts would classify the items and also to identify any potential problems with categorisation process (Sabbag et al., 2018). It was important to ensure that each item in the test assessed a single SL skill, that is, one of the four defined skills: interpreting, communicating, evaluating and decision-making. The three colleagues worked independently and were asked to fill out a skill review form (see Appendix B). The form consists of three sections: the definition of the four skills as a reference for the validators, a table of items for which each validator had to predict the assessed skills and each validator's choice of which item should be included in the test. By including the definitions of the four skills, the review form was used to reduce the personal biases of validators (Sabbag et al., 2018). In addition, the 14 items were attached separately; they were randomly ordered. The validator's prediction of the skill assessed by each item was then documented by a researcher for cross-checking, and the validator's item choice was recorded as an additional consideration.

Cross-checking between the researcher's listed skill and each validator's review was undertaken in two stages, to reach an absolute agreement. The researcher conducted the first stage independently, comparing the skills listed by the researcher to each validator's predicted skills (see Table 4.7). This stage focused on the disagreement while considering the validator's reasoning for their prediction. The second stage was a closed discussion that served as a hearing session on the two parties' basic reasoning. Among the three validators, the disagreement of the predicted skill was discovered in validator 1. The researcher and validator 1—through a closed discussion—were able to reach an agreement on the skills assessed by all items. This cross-checking contributed to the instrument's validity by clarifying that each item assessed one of the four skills.

Table 4.7

First Stage of Validating the Items' Assessment of Skills

Context	Assessed skill			
	Researcher	*Validator 1*	*Validator 2*	*Validator 3*
The Production Mean	I	C	I	I
The Most Production	I	I	I	I
Faulty Electronics	E	E	E	E
YouTube Viewers	C	C	C	C
In Which Month?	I	I	I	I
Mathematics Scores	E	E	E	E
Domestic Waste	C	I	C	C
The 100-Metre Race	D	D	D	D
The Employees	E	I	E	E
How Many Students?	I	E	I	I
Dramatic Decline	E	E	E	E
Which Motorcycle?	D	D	D	D
Average Size	E	E	E	E
More Stock	D	D	D	D

Note. I = interpreting; C = communicating; E = evaluating; D = decision-making.

4.2.2 Results of Cognitive Interview Protocol Development

Conducting a cognitive interview with students has two purposes: contributing to the validity of the assessment items and validating the students' thought processes (Desimone & Le Floch, 2004). A cognitive interview collects additional verbal information about how the participants responded to the items (Beatty & Willis, 2007; Conrad & Blair, 1996; Desimone & Le Floch, 2004). The cognitive interview involves four problem-solving processes: 'comprehending the item', 'retrieving relevant information', 'making a judgement based upon the recall of knowledge' and 'mapping the answer onto the reporting system' (Desimone & Le Floch, 2004). Technically, this interview was conducted in semi-structured manner to facilitate flexible exploration during the interview (Magaldi & Berler, 2020). The flexibility to delve deeper into students' responses and ask clarifying questions throughout the four problem-solving process eventually improved the depth and richness of the information gathered. This interview was conducted after the test and conjectures about the students' reasoning were made after the test based on their written responses. As a result, the interview protocol was developed to validate the conjectures and reveal cognitive processes not described in the students' written responses.

The initial version of the cognitive interview protocol was developed in parallel with the development of the assessment items. The interview protocol consisted of two parts: the interview technique and the interview script (see Appendix C). The interview technique covered the definition of an SL interview and described the preparation and interview setting. The interview script included four problem-solving processes as described above. The interview questions were developed, along with appraisal statements responding to the students' answers, to cover these four problem-solving processes. For example, *Please read the question aloud!* was the first instruction asked of a student to indicate his or her item comprehension, *When you give your answer, what are you thinking?* was a probing question seeking the students' stages in solving the problem, and *Your reaction is helpful; thank you.* was an encouraging response to students' answers. Table 4.8 presents the initial script for the cognitive interview.

Table 4.8

Initial Interview Script

Step	Script
Beginning the interview	*Before we begin this interview, let's have a proper introduction. My name is Badrun, and you are* [insert interviewee's name]. *It's nice to meet you.*
(4 minutes)	*First of all, I would like to thank you very much for participating in this interview. Your participation will help me understand what is going on in your mind while you are working with data-based problems. There will be no right or wrong answers; therefore, you don't have to be afraid.*

Table 4.8 (continued)

Step	Script
	Any information you provide during this interview will be recorded. Is that OK, [insert interviewee's name]*? Thank you.* *During this interview, please think aloud as you're solving the problems. That means say anything—whatever you think; I'm interested in hearing all your thoughts and reactions.* (**Repeat and emphasise this information.**) *We will now begin this cognitive interview at time* [insert time].
Conducting the interview	Comprehension of the item: *Please read the question aloud!*
(25 minutes)	**Probe:** *Do you understand what you've just read?*
	If no, probe: *Which particular information in this item was difficult to understand?*
	If yes, probe: *Can you explain it in brief?*
	That's great. Thinking out loud like this is just what I need.
	Retrieval of relevant information: *I am interested in what you are thinking as you retrieve relevant information from the problem; do whatever you need to help you think aloud.*
	Probe: *Why do you think that might be the relevant information?*
	Thank you—your responses are really helpful.
	Judgement-making based upon the recall of knowledge: *Do you understand what this question is asking?*
	Probe: *Then, what do you need to do to solve it?*
	That's fine; you are talking through your reaction, and it is very helpful for me.
	The process of mapping the answer onto the reporting system: *Please start explaining the answer you've written for this question by saying it out loud.*
	Probe: *When you gave your answer, what were you thinking?*
	Your reaction is helpful; thank you.
	(This set of questions applies to all test items)
Closing the interview	*Thank you for taking the time to participate in this interview. If you have any comments to share, please feel free.*
(1 minute)	*That concludes this interview, and I will now stop the recording at time* [insert time].

Note. This interview script has been translated into English.

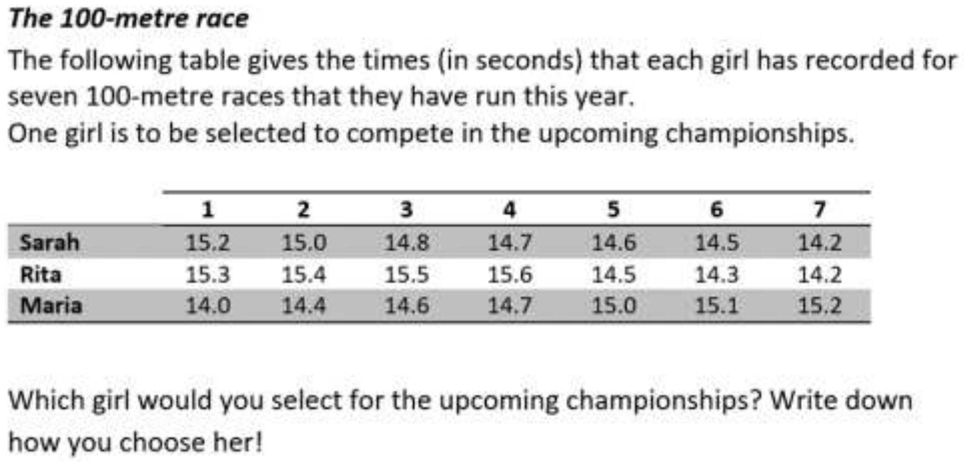
The list of questions in Table 4.8 is the general version and needed to be adapted to be specific to each item. Further, the probing questions are examples and can be expanded in response to the student’s answers. This initial version of the interview protocol was later piloted for refinements and to provide the interviewer (i.e., researcher) with enough experience to conduct an SL cognitive interview. It was expected that the quality of the questions and the interviewer’s experience would enable him to reveal the cognitive processes of the interviewed students.

4.2.3 Results of Descriptor Development

Recall that the levels employed to characterise students’ SL levels were the six hierarchical levels whose descriptors cover the three SL components (see Section 2.6 in Chapter 2). As each item assesses a specific SL skill that incorporates three SL components, the descriptors for each item were developed with reference to those components. To illustrate how these descriptors were developed, The 100-Metre Race item will now be used as an example (see Figure 4.1). The features of this item are a table containing recorded times for three runners and surrounding text that explains the context. The question is displayed below the table and is designed to assess the student’s decision-making skill by asking them to choose one runner of the three to compete in the upcoming championship race.

Figure 4.1

The 100-Metre Race Item



The 100-metre race

The following table gives the times (in seconds) that each girl has recorded for seven 100-metre races that they have run this year.

One girl is to be selected to compete in the upcoming championships.

	1	2	3	4	5	6	7
Sarah	15.2	15.0	14.8	14.7	14.6	14.5	14.2
Rita	15.3	15.4	15.5	15.6	14.5	14.3	14.2
Maria	14.0	14.4	14.6	14.7	15.0	15.1	15.2

Which girl would you select for the upcoming championships? Write down how you choose her!

The component-based descriptors theoretically developed across the six levels in Chapter 2 (Table 2.8) were transferred into the item's level descriptors for each of the three components (these descriptors were then termed the component-based item descriptor). This process was enriched by the expected student responses provided by Sharma et al. (2012) and the teacher's guide of the Indonesian mathematics textbook from which this item was developed (As'ari et al., 2017b, p. 352). Based on these sources, students' responses to The 100-Metre Race item were predicted for each of the six levels (idiosyncratic, informal, inconsistent, consistent non-critical, critical and critical mathematical). For example, the idiosyncratic level in the general descriptors has the keywords 'personal engagement with the context' (for text and context), 'read specific values' (for representation) and 'perform simple calculation' (for statistical-mathematical knowledge). Consistent with the expected student responses provided by Sharma et al. (2012) and the textbook, the idiosyncratic level descriptor for this item predicts that students will use personal experience in interpreting the context of running (text and context), choose inappropriate data from the table (representation) and perform calculations based on numbers in the table that do not relate to the question (statistical-mathematical knowledge). This process was applied for all six levels and resulted in the component-based item descriptor shown in Table 4.9.

Table 4.9

Development of the Level Descriptors for The 100-Metre Race Item

Level	Component-based descriptors	Initial component-based item descriptor
Idiosyncratic	*Text and context*: Students demonstrate non-existent or personal engagement with contexts. *Representation*: Students show that their personal beliefs and experiences underlie their basic graph and table reading (e.g., reading cell values). *Statistical-mathematical*: Students guess the answer, do one-to-one counting, pick a random value, perform simple calculations, select the largest number or take other unreasonable steps.	*Text and context*: Students interpret the context of running using personal experience and use it to choose one of the runners. *Representation*: Students choose inappropriate data from the table. *Statistical-mathematical knowledge:* Students make calculations based on the numbers in the table but not relating them to the question.
Informal	*Text and context*: Students show engagement with contexts, but their engagement is colloquial or informal (reflecting intuitive or non-statistical beliefs) and reflects irrelevant aspects of the context. *Representation*: Students may be successful at some of the more basic table or graph reading, such as comparing cells to determine the highest or most frequent value and identifying the smallest data value. *Statistical-mathematical*: Students perform basic one-step table and graph calculations (such as addition and subtraction) based on the values observed, but sometimes accompany the calculation with an imaginative story.	*Text and context*: Students interpret the context of running using everyday experience (the winner has the biggest number) and use it to choose one of the runners. *Representation*: Students use the data for three runners in seven races, but do not interpret the values as times. *Statistical-mathematical knowledge:* Students use the total of the data to choose the best runner (the one with the longest total time).

Table 4.9 (continued)

Level	Component-based descriptors	Initial component-based item descriptor
Inconsistent	*Text and context*: Students demonstrate selective or inconsistent engagement with contexts (depending, to some extent, on the format of the items). *Representation*: Students tend to interpret the graphical or tabular details rather than the context of the graph or table and fail to describe the relationship between data. *Statistical-mathematical*: Students make conclusions, but those conclusions are not always accompanied by suitable statistical or mathematical justifications.	*Text and context*: Students understand partially the context of a running competition but still use an informal interpretation. *Representation*: Students read the data in a table that displays the times for three runners over seven races but fail to recognise the relationships between the data. *Statistical-mathematical knowledge:* Students use the mean or mode to figure out who is the best runner but inappropriately choose the largest mean or the longest time.
Consistent non-critical	*Text and context*: Students show appropriate engagement with contexts, but often do so in a non-critical manner. *Representation*: Students make sense of the data presented in a graph or table while partially recognising the context, focus on a single relevant aspect of the data or compare the data within the table or graph. *Statistical-mathematical*: Students accurately and appropriately use simple statistical and mathematical concepts, including those associated with graph characteristics.	*Text and context*: Students appropriately understand the context of a running competition in which the winner is the one with the shortest time. *Representation*: Students read the data in a table that displays the times of three runners in seven races and recognise the relationship between them. *Statistical-mathematical knowledge:* Students use the mean to choose two runners with the same average.
Critical	*Text and context*: Students demonstrate critical engagement with familiar contexts and less-critical engagement with unfamiliar contexts.	*Text and context*: Students critically understand the context of a running competition. *Representation*: Students recognise that the best runner is the one with the lowest mean.

Table 4.9 (continued)

Level	Component-based descriptors	Initial component-based item descriptor
	Representation: Students demonstrate awareness of the relevant features of a graph or table and awareness of the integration of more than one relevant aspect of data to show a relationship. *Statistical-mathematical*: Students demonstrate qualitative interpretation and sophisticated use of mathematical or statistical concepts.	*Statistical-mathematical knowledge:* Students choose one of two runners with the same average and justify that choice by, for example, choosing the one who won most often.
Critical mathematical	*Text and context*: Students demonstrate critical, questioning engagement with familiar and unfamiliar contexts. *Representation*: Students critically summarise the association between the variables shown in a graph or table and relate it to the context. *Statistical-mathematical*: Students perform sophisticated or critical statistical and mathematical tasks associated with statistical concepts such as central tendency and dispersion measures.	*Text and context*: Students understand critically the context of running competitions and factors to be considered in choosing the best runner. *Representation*: Students summarise the data in the table (such as mean, variation and trend) to select the best runner. *Statistical-mathematical knowledge:* Students choose one out of two runners with the same average and justify the choice, for example, by comparing the trend.

The process of developing component-based item descriptors for The 100-Metre Race item applied similarly to the other 13 items. Further refinement was applied based on the empirical data obtained from the pilot test and pilot interviews; this will be described in Section 4.3.3.

4.3 Piloting and Refinement

4.3.1 Results of Piloting and Refining Test Items

In this section, the process and results of the three-stage piloting are presented. The piloting was conducted from August to September 2019. These stages were Pilot Interview I, the Pilot Test and Pilot Interview II. Table 4.10 summarises those stages and the results obtained at each of them. To provide a comprehensive understanding, the results of each stage are explained in the subsequent sections, along with the assessment item refinement.

Table 4.10

Piloting Stages and Output

Stage	Description	Output
Pilot Interview I	Examine the clarity of the items and the cognitive processes of four students with average academic achievement.	The revised 14 items.
Pilot Test	Trial the items with 12 low- to high-achieving students to determine the test administration process and duration and discover how students will respond to the items.	Test administration procedure, test duration and a sample of student written answers.
Pilot Interview II	Investigate the cognitive processes underlying students' written pilot test answers to enhance the quality of the interview protocol.	Ten revised items along with their interview protocols and level descriptors.

4.3.1.1 Pilot Interview I

In Pilot Interview I, the 14 statistical items developed prior to the piloting were used to search for potential insights from the students' perspectives. This interview aimed to investigate students' understandings of each item's features: the texts providing context for the problems, the graph or table displaying the context-related data, the question assessing one of the skills and the statistical and mathematical concepts needed to solve the problem. Understanding the problem from the students' perspective could provide evidence of the item's validity and minimise the incorrect answers caused by instrument errors (Ziegler & Garfield, 2018). In other words, the pilot interview provides evidence of response process validity in terms of how students respond to the items (Sabbag et al., 2018).

With the support of their mathematics teachers, two Year 9 and two Year 12 students with an average level of knowledge voluntarily participated in Pilot Interview I. This interview was conducted at the beginning of the first semester (August) in 2019. The Year 9 students were from a public *sekolah* (a school under MoEC-RT), while the Year 12 students were from a private *madrasah* (an Islamic school under MoRA). All these schools were from Jombang (one of the cities in which the actual data collection would take place). It was expected that the students' responses would illustrate the typical thought process and responses of Year 9 and 12 students with average levels of knowledge. Moreover, this empirical insight into the students' responses was expected to enrich further the items that had previously been theoretically developed. In addition, students' responses proved crucial in helping to confirm and evaluate the previous conjectures about students' cognitive processes.

To a certain extent, this semi-structured interview was also intended to test the interview protocol. Before the interview started, the students were made comfortable by briefly informing them how the interview would be conducted (Reinhart et al., 2022). Knowing they were unfamiliar with cognitive interviews, the interviewer (i.e., the researcher) asked students to please think aloud as they comprehended items and mapped the answers. That meant students could say anything in the process of understanding and answering the problems. Items 1 to 14 were discussed with each of the Year 9 students separately. While interviewing, the interviewer took notes to record unanticipated actions. This information helped the interviewer improve his practice when interviewing the two Year 12 students in the following days.

The location and timing of Pilot Interview I depended entirely on the students' availability. They were interviewed individually in separate locations and at different times. Each student participated in an approximately one-hour interview. The interviewer started by asking them to read the problem aloud to check their item comprehension. The students' think aloud responses and the interviewer's follow-up probes helped identify the cognitive processes of students attempting to comprehend each item, retrieve relevant information, recall related knowledge and map strategies to answer the questions. These interviews were audio-recorded, with the students' consent, for further review and item refinement (see the consent form for parents in Appendix D).

After finishing Pilot Interview I, the interviewer began searching for possible clues for assessment items and interview protocol refinement by listening to the audio recording; however, note that interview protocol refinement is presented separately (see Section 4.3.2), this section only focuses on item refinement. The researcher conducted this work with the help of one research assistant. Both the research assistant and researcher identified some clues from the Pilot Interview I that were worthwhile for item refinement, including the wording in the text, the graphical and tabular displays and the responses. Table 4.11 shows some excerpts from the students' responses to The 100-Metre Race item that further motivated the refinements.

Table 4.11

Pilot Interview I Results for The 100-Metre Race Item

Feature	Students' comprehension	
	Year 9	*Year 12*
Text	Students read the texts and directly related the information to the table.	Students read the texts carefully and restated the text to express their understanding.
Table	Understanding: *One of them in this table must be the best.* *So, this is the time for each student* [pointing through the recorded times]. *So, this is the time, 15.2 seconds, and 1 to 7 is the round.*	Understanding: *The numbers in the tables show the times.* *Sarah in the first race is 15.2 secs, Rita 15.3 secs, Maria 14.0 secs, and Maria is the quickest because* [she] *only* [took] *14.0 seconds.* Misunderstanding: *What do these numbers mean? 14.2?*
Statistical-mathematical knowledge	The most stable runner *I select Maria, oh, Rita, because she was the most stable from the first to the fifth race.* The biggest time *The winner in the seventh race is absolutely Maria and in the second race is Rita.* The fastest mean *So, find the mean first and select the one with the shortest mean.*	The fastest mean *Looking for the shortest time by finding the mean* *Finding the mean of each and selecting the quickest* *The mean is finding the total time and dividing by how many races.*

Students' oral responses to The 100-Metre Race item revealed both students' understandings and challenges. Students showed contextual-tabular understanding, such as '*so, this is the time, 15.2 seconds, and 1 to 7 is the round*' and '*Sarah in the first race is 15.2 secs, Rita 15.3 secs and Maria 14.0 secs, and Maria is the quickest because* [she] *only* [took] *14.0 seconds*'. Their responses indicated their understanding of the context of seven races and the recorded time over seven races. However, some challenges were also revealed. These challenges were further checked to determine whether they were caused by the students' lack of knowledge (participant variability) or the items' wording ambiguity (instrument error). If students' challenges were caused by their limited knowledge, the items did not necessarily need to be refined—otherwise, the items needed to be revised to prevent confusion. As presented in Table 4.11, the indication of students' challenges could be identified in their responses such as '*Rita, because she was the most stable from first to fifth race*' and '*The winner in the seventh race is absolutely Maria and in the second race is Rita*'. These responses indicated the student's lack of contextual-tabular understanding; thus, no revision was needed.

After analysing the students' interviews for all 14 items, three types of refinements were made (see Appendix E for the revised items). The first type was applied within the contextual texts. These changes varied from merely adding a comma (punctuation) to changing or adding words and numbers. For example, a comma was added to The Employees item to prevent students from reading the whole sentence in one breath, which could lead to a contextual misunderstanding. Another example was changing the criteria of selection in the Which Motorcycle? item (from 2010 to 2011) to better align with the year production data in the table. Three items needed changes to the phrasing: Faulty Electronic, Mathematics Scores and The 100-Metre Race. For example, the word '*pada*' ('on' in English) was used in the Mathematics Scores item instead of '*dalam*' ('in' in English)' for the translation of 'in' in the phrase '*in this test*'. Finally, some sentences were created to provide contextual background for Domestic Waste, How Many Students? and Dramatic Decline items.

The second type of amendment relates to the items' graphical or tabular displays. These changes were applied to four items: Faulty Electronic, Which Motorcycle? How Many Students? and Dramatic Decline. For the Faulty Electronic and Which Motorcycle? items, minor changes were only applied to one column caption in the table to provide a more interactive table. For How Many Students? and Dramatic Decline items (they used the same context and graph), the adjustment was applied to one bar in the graph and the x-axis. More explicitly, one bar in the graph informing the number of Year 6 students and a zigzag line on the x-axis indicating that the data displayed did not include the number of students below Year 6 were omitted as they had no importance.

Finally, the third type of item refinement was the question's wording. This edit was applied to six items. The revisions varied from adding or omitting words to modifying the question. Words were changed to the Indonesian version of The 100-Metre Race item, while words were omitted from The Most Production item. In addition, some instructions were modified for YouTube Viewers, In Which Month, How Many Students? and Dramatic Decline items. The refinements of four items are exemplified in Table 4.12, while the full version of the revised items is available in Appendix E.

Table 4.12

Assessment Item Refinement after Pilot Interview I

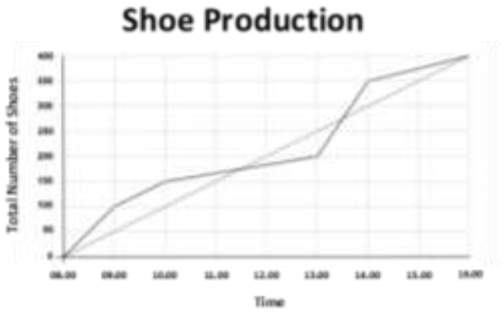
Results of Pilot Interview I	Revised item after Pilot Interview I
The context, graph and question all remained the same.	Interpreting item (I)  The solid line (—) on the graph shows the number of shoes produced by a home industry during a particular day. The dotted line (- -) shows what the total number of shoes produced would be if the rate of production were constant. What was the mean number of shoes produced per hour? Explain how you got it.

Table 4.12 (continued)

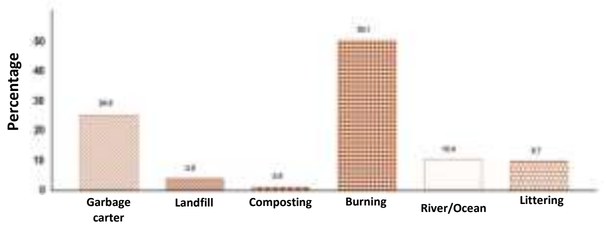
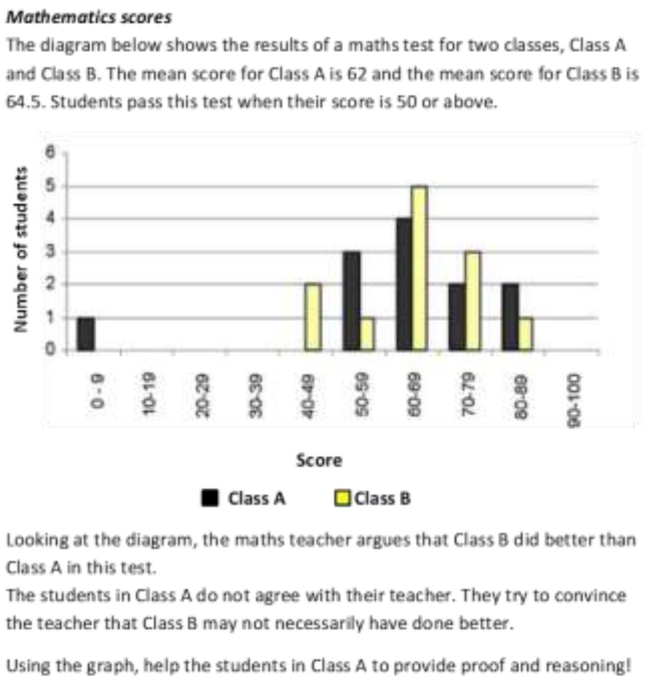
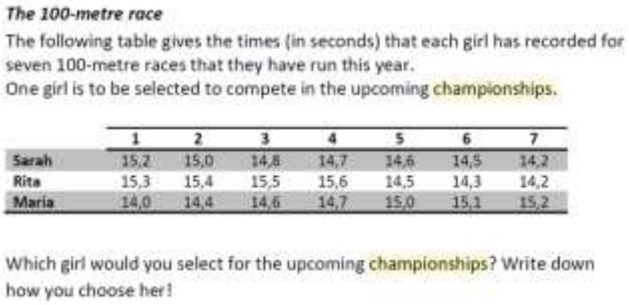
Results of Pilot Interview I	Revised item after Pilot Interview I
The context: One sentence was added before the bar graph instead of the bar title to match it with other items. That sentence explains the context for the bar graph, that is, the various actions undertaken by Indonesians concerning their domestic waste. The graph: The graph convention was maintained. The question: The graph convention was maintained.	Communicating item (C) ***Domestic waste*** The bar graph below shows the various actions taken by Indonesians towards household waste in 2013.  To make your friend informed, summarise the important information from the graph about the Indonesian people's awareness of domestic waste management!
The context, graph and question all remained the same.	Evaluating item (E)  ***Mathematics scores*** The diagram below shows the results of a maths test for two classes, Class A and Class B. The mean score for Class A is 62 and the mean score for Class B is 64.5. Students pass this test when their score is 50 or above. Number of students 0 1 2 3 4 5 6 0 - 9 10-19 20-29 30-39 40-49 50-59 60-69 70-79 80-89 90-100 Score ■ Class A □ Class B Looking at the diagram, the maths teacher argues that Class B did better than Class A in this test. The students in Class A do not agree with their teacher. They try to convince the teacher that Class B may not necessarily have done better. Using the graph, help the students in Class A to provide proof and reasoning!

Table 4.12 (continued)

<table>
<tr><th>Results of Pilot Interview I</th><th>Revised item after Pilot Interview I</th></tr>
<tr><td>The context, graph and question all remained the same. However, a small change was applied to the translation of ‘championships’: from <i>pertandingan</i> to <i>perlombaan</i>.</td><td>Decision-making item (D)

<i><b>The 100-metre race</b></i>
The following table gives the times (in seconds) that each girl has recorded for seven 100-metre races that they have run this year.
One girl is to be selected to compete in the upcoming championships.
<table><tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr><tr><td>Sarah</td><td>15,2</td><td>15,0</td><td>14,8</td><td>14,7</td><td>14,6</td><td>14,5</td><td>14,2</td></tr><tr><td>Rita</td><td>15,3</td><td>15,4</td><td>15,5</td><td>15,6</td><td>14,5</td><td>14,3</td><td>14,2</td></tr><tr><td>Maria</td><td>14,0</td><td>14,4</td><td>14,6</td><td>14,7</td><td>15,0</td><td>15,1</td><td>15,2</td></tr></table>
Which girl would you select for the upcoming championships? Write down how you choose her!</td></tr>
</table>

4.3.1.2 Pilot Test

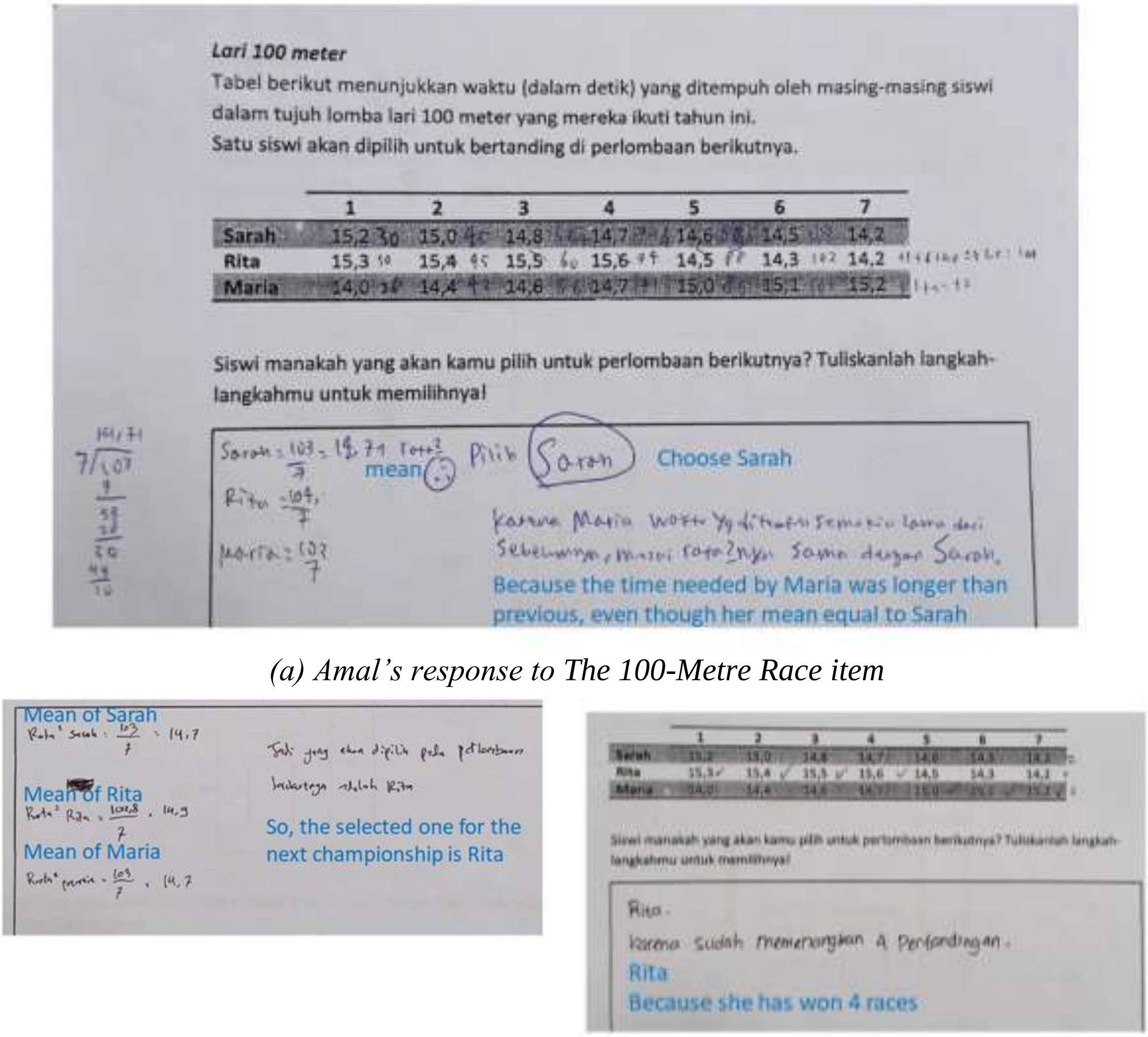
Following Pilot Interview I, a pilot test, which was a small-scale trial, was conducted with representatives of the actual study participants. In this pilot test, 12 students (six Year 9 and six Year 12 students) were selected by their mathematics teachers among those who agreed to voluntarily participate. The Year 9 students were from a public *madrasah*, while the Year 12 students came from a public *sekolah*. Further, these students represented three levels of knowledge in mathematics: high, medium and low levels (from their mathematics teachers’ perspectives). The intent behind involving these students was to collect a wide range of responses from students at various knowledge levels. The wider the range of responses emerging from students, the better the pilot test would predict the spectrum of students’ responses in the actual test.

The pilot test informed the feasibility of administering the actual test in terms of the test duration and the testing process, in addition to aiding the understanding of students’ written responses. The pilot test used the 14 items refined after Pilot Interview I and lasted about 105 minutes for Year 12 students and about 120 minutes for Year 9 students. However, in both grade levels, some students completed the test in the first 90 minutes. The students’ times to completion were compelling evidence that the time required per item was seven to nine minutes, which was ideal. Regarding the testing process, most students solved all items independently, though some requested clarification of specific wording in items' contexts or questions. These requests were noted and investigated to determine whether they were caused by ambiguity in the item's wording or students' lack of knowledge. Further, these notes were considered during later item refinement, supplementing what would be found in Pilot Interview II.

Figure 4.2 shows three examples of the students' responses, showing variation in the students' understandings and challenges of The 100-Metre Race item. The students were Amal, Zainuddin and Oktav (all names used in this study are pseudonyms). From Amal's responses, it can be inferred that he demonstrated a critical understanding of this problem by applying two selection methods. From his written answer, it was conjectured that he started working by calculating the total time each runner clocked for the seven races. The calculation method he applied indicated his sophisticated number sense because he added only the integers first and added the decimal numbers later. Once he found the total time, he continued calculating the mean time for each runner. Finding that there were two runners with the same mean, Amal used the trend to decide further which runner should compete in the next championship. This thinking eventually led Amal to correctly choose Sarah from the three runners.

Figure 4.2

Three Examples of Students' Responses during Pilot Test



Lari 100 meter

Tabel berikut menunjukkan waktu (dalam detik) yang ditempuh oleh masing-masing siswi dalam tujuh lomba lari 100 meter yang mereka ikuti tahun ini.
Satu siswi akan dipilih untuk bertanding di perlombaan berikutnya.

	1	2	3	4	5	6	7
Sarah	15,2	15,0	14,8	14,7	14,6	14,5	14,2
Rita	15,3	15,4	15,5	15,6	14,5	14,3	14,2
Maria	14,0	14,4	14,6	14,7	15,0	15,1	15,2

Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!

Sarah = $\frac{103}{7}$ = 14,71 rata2 mean Pilih Sarah Choose Sarah

Rita = $\frac{104}{7}$

Maria = $\frac{103}{7}$

Karena Maria waktu yg ditempuh semakin lama dari sebelumnya, meski rata2nya sama dengan Sarah.
Because the time needed by Maria was longer than previous, even though her mean equal to Sarah

(a) Amal's response to The 100-Metre Race item

Mean of Sarah
Rata² Sarah = $\frac{103}{7}$ = 14,7

Mean of Rita
Rata² Rita = $\frac{104,8}{7}$ = 14,9

Mean of Maria
Rata² Maria = $\frac{103}{7}$ = 14,7

Jadi yang akan dipilih pada perlombaan berikutnya adalah Rita
So, the selected one for the next championship is Rita

(b) Zainuddin's response to The 100-Metre Race item

	1	2	3	4	5	6	7
Sarah	15,2	15,0	14,8	14,7	14,6	14,5	14,2
Rita	15,3 ✓	15,4 ✓	15,5 ✓	15,6 ✓	14,5	14,3	14,2
Maria	14,0	14,4	14,6	14,7	15,0 ✓	15,1 ✓	15,2 ✓

Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!

Rita.
Karena sudah memenangkan 4 Pertandingan.
Rita
Because she has won 4 races

(b) Zainuddin's response to The 100-Metre Race item *(c) Oktav's response to The 100-Metre Race item*

From the conjectures made about Amal's thought process, his written answer was then attributed to the three components. This attribution was beneficial for developing the protocol for interviewing him. In terms of text and context, he understood the rules of a running competition and that three runners had competed in seven races, and that one best runner must be chosen from them. In terms of representation, he was able to see that the table consisted of the times recorded by three runners to reach the finish line in seven races. In addition, he could relate these numbers to the context, in which the shorter the time, the quicker the runner reaches the finish line. In terms of statistical-mathematical knowledge, he applied mean and trend as selection methods, showing sophisticated number sense.

Like Amal, Zainuddin was able to determine the total time and mean time of each runner. However, he displayed a contextual misunderstanding when choosing the best runner. According to him, the best runner was the one whose mean was the highest. As a result, he selected Rita instead of Sarah or Maria because Rita had the highest mean. Consequently, in terms of text and context, his conjectured response suggested that he understood that three runners were competing in seven races and one best runner had to be chosen. However, he incorrectly related the winner in the running competition to other competitions in which the largest number determines the winner. In terms of representation, he was able to see that the table consisted of times taken by three runners to reach the finish line in seven races. However, his interpretation was incorrect, influenced by his incorrect contextual understanding. In terms of statistical-mathematical knowledge, he applied the mean as a selection method and could perform the correct calculation for the total time and mean.

Oktav's response provides a further comparison. Oktav used mode instead of mean to select the best among the three runners. Her ticks in the table indicate her thought processes and selection technique. She chose the runner with the highest number as the winner in each race (in fact, the numbers are the times in seconds). This technique led Oktav to choose Rita, thinking she had won (was ticked in) four out of seven races. Based on this logic, Oktav's conjectured response was analysed in terms of the three SL components and later used as a reference for the interview. In terms of text and context, Oktav seemed to understand that there were seven races and found challenges to understand that the numbers in the table indicated each runner's time. In terms of representation, Oktav was able to read the data in the table that presents the times for the three runners over seven races but failed to relate it to the context that the winner should be the one with the shortest time. In terms of statistical-mathematical knowledge, Oktav applied the mode to choose the best runner, but her inappropriate contextual understanding of the winner resulted in the wrong choice.

In summary, the findings from the pilot test informed the test duration, the testing process and the conjectures about students' written responses. Regarding the test duration, the time needed by students in both grade levels was seven to nine minutes per item. Regarding the testing process, the test went according to plan, but there were nevertheless students who asked for clarification of certain wording in the problems during the test. This need for clarification was investigated to determine whether it was caused by wording ambiguity. Finally, the responses of the 12 students were subjected to a thorough analysis. This process led the researcher to conjecture about what the students did and did not understand about the three SL components, which was useful for Pilot Interview II.

4.3.1.3 Pilot Interview II

As the final stage of piloting, Pilot Interview II examined the cognitive processes underlying the written responses of the 12 students who took the pilot test. This interview aimed to verify that the students had comprehended the items and enhance the quality of the interview protocol to be used for the actual data collection. Before the interview, each written response from the 12 tested students was examined to understand their trajectory from item comprehension to the answering process, including what they did and did not understand. Afterwards, a unique cognitive interview protocol was developed to delve as deeply as possible into each student's cognitive processes. The researcher expected that by using this interview protocol, the students' trajectory from item comprehension to the answering process would be traced and identified. The interviews were conducted across four days. Each of the 12 students was interviewed, except for one Year 9 student who was sick during the interview schedule and could not be rescheduled. Each interview was audio-recorded, and at the end of each day, a reflection was performed to improve the interview protocol for the subsequent days. After many of these reflections, a small number of the probing questions were modified to elicit more detailed responses.

The results of the interviews were later analysed, focusing on the students’ responses and comprehension of the items as indicators of the clarity of the item’s wording, sentence structure, representation and questions. The interview results of three of the 12 students (the three mentioned in the pilot test section above) were analysed to show their different levels of understanding. The following analysis of students’ interviews are based on three written responses from Amal, Zainuddin and Oktav as previously presented in Figure 4.2.

4.3.1.3.1 Amal’s Understanding

Amal, a Year 12 student, showed critical understanding in his written response. Amal made sense of the context, a running competition, and the data series expressed in the table. Based on his written response, his technique to select one best runner to compete in the next championship revealed his expertise in number sense and his conceptual understanding of data measures of centre and trend. It was conjectured that he first tried selecting the best runner using the mean and then employed the trend once he realised that Sarah and Rita had the same mean. The interview confirmed his critical understanding but revealed that the steps he used to solve the problem were not as conjectured.

When asked to explain his process of item comprehension, Amal appeared to incorporate the information in the texts and the data in the table. When asked to explain what he understood from the texts above the table, he replied, ‘*I just read, and there was no crucial information in the texts*’. Although Amal (A) made this statement to the interviewer (I), it did not indicate that he had no understanding of the text because he then looked directly at the table and gave a powerful and comprehensive interpretation:

A: *1, 2, 3, and 4 is like how many competitions, and the numbers* [pointing to the times] *are the recorded values for 100 metres; the finish [time is] 15.2 seconds, 15.3 and 14.0* [pointing to Sarah, Rita and Maria's times in the first race].

I: *What about the calculation you performed on the right of the table? Can you please explain it?*

A: *The calculation was actually performed at the end, not in the beginning.*

I: *Then please explain what you were doing in the beginning.*

A: *First, I observed the numbers, the finishing times from 1 to 7; it is a running race, so it should be the fastest completion time. I observed that Sarah, the more [she raced], the faster [her finishing time]. Rita was [the] same [as] Maria. Maria, the more [she raced], the slower [her finishing time]. After that, [I] made [the] hypothesis that the answer is Sarah. Then, to prove the hypothesis, [I] calculated the total time for all students and [the] mean for Sarah. Sarah and Maria were [the] same, and Rita's mean [was] bigger. Although Maria and Sarah were [the] same, Maria, the more [she raced], the slower [her finishing time], so I chose Sarah.*

His explanation revealed that he started by selecting one best runner from observing the trends in the three runners' finishing times over seven races and then hypothesised about who the best runner was. He hypothesised that Sarah was the best runner as her finishing time improved over the seven races. To prove his hypothesis that Sarah was the best candidate to compete, Amal calculated the mean for all three runners for comparison. When he discovered that Sarah's mean was both the smallest and similar to Maria's, he took a step back to compare their trends. Finally, Amal chose Sarah because of her trend for 'getting faster', whereas Maria displayed a 'getting slower' trend.

Through his strategy of calculating the mean for three runners, Amal showed an excellent understanding of number sense. When Amal was asked to explain his calculation, he answered:

> *In calculating, I did not like it if the total was not an integer; I didn't consider the digit after [the] decimal point. That was calculated at the end as [for Maria] 0.4 + 0.6 = 1, 0.7 + 0.1 + 0.2 = 1, and Rita's was the most difficult to count.*

Amal's calculation can be interpreted from this explanation as beginning with the integers side by side. In the case of Maria (see again Figure 4.2), the integer for the first race (14) was added to the integer for the second race (14), and the result (28) was written between those integers. Then, 28 was added to the integer for the third race (14), giving 42, written between the integers for the second and third races. These additions continued until the last race, resulting in a total of 101 for the seven races. Then, Amal added all the decimal fractions. The decimal fractions for the second and third races (0.4 and 0.6) were added to make 1, and the decimal fractions for the fourth, sixth and seventh races (0.7, 0.1 and 0.2) were added to make 1. Therefore, the total time for Maria was 101 + 1 + 1 = 103. This number sense–based calculation was also applied to the other two runners.

In conclusion, Amal's critical responses showed that the features of The 100-Metre Race item had proved comprehensible. First, Amal's comprehension of the context through reading the texts and interpreting the data in the table indicated that these features were not ambiguous to him. Further, the question assessing the decision-making skill was also correctly addressed by him. His utilisation of two means of selection, mean and trend, proved that such a question could prompt students to demonstrate critical and comprehensive statistical and mathematical knowledge using data from the table.

4.3.1.3.2 Zainuddin's Challenges

Zainuddin, a Year 9 student, showed contextual challenges in interpreting the times the three runners took to reach the finish line. Although his calculation of the mean time each runner took was correct, his contextual understanding failed him. During the interview, he said that he started trying to comprehend the items from the table, the context and then continued reading the question. Zainuddin (Z) described his thought process to the interviewer (I) as follows:

I: *From which part did you start trying to comprehend the item?*

Z: *From the table.*

I: *What information did you get from the table?*

Z: *There were three students who competed in running seven times and got different results.*

From his explanation, it seems Zainuddin had no problem understanding the context or correctly using it to interpret the data represented in the table. He further showed his understanding of statistical and mathematical knowledge by using the mean as a method of selection:

Z: *I added up all the times from race 1 to race 7.*

I: *Do you mean for each student?*

Z: *Yes, for each student.*

I: *After the total was found, what was next?*

Z: *Divided by seven because there were seven competitions.*

I: *What did you want to find?*

Z: *To find the mean.*

At this point, all his arguments still made sense as he used the mean as a selection method. The total time for each runner was correct, and the resulting means were also correct: Sarah, 14.7; Rita, 14.9; and Maria, 14.7. However, the challenges he faced was revealed when he began to explain the selection process.

R: *What did you think of after finding the mean for those three runners?*

Z: *I immediately thought that Rita should compete in the next championship because she [had] the highest mean.*

R: Then, what happens to both Sarah and Maria, who have the same mean?

Z: *They both do not join the competition.*

From his explanation, it was revealed that he had a challenge in understanding how long it took the three runners to cross the finish line. He thought that the higher the mean, the better the runner, not the reverse.

To this end, it could be concluded that Zainuddin could understand from the texts and table that this item was about three runners who completed seven races and that their times to reach the finish line were recorded in the table. In addition, he also correctly understood that the question assessed his decision-making skill. Zainuddin's choices to use the mean as a selection method and to apply the mean formula did not show any signs of error. Unfortunately, however, he falsely interpreted the means he calculated for the three runners by choosing the highest as the best. Thus, the challenges Zainuddin faced was purely caused by his lack of knowledge and not by the item's ambiguity.

4.3.1.3.3 Oktav's Challenges

Oktav (O), a Year 9 student, started trying to comprehend the item from the question first and then continued on to the text and table. She realised that the context was about three runners running seven 100-metre races. However, she could not recognise the numbers as the times each runner took to run 100 metres; instead, she saw them as scores. She concluded that the winner in each race was the one with the highest score. She ticked this 'winner' in each race and eventually selected the runner who had ‘won’ the most races to compete in the next championship. The challenges she faced was reflected in the interview below:

I: *From the table and the text, what can you understand?*

O: *Here is Rita* [pointing to Rita’s time], *who is the highest.*

I: *If I ask you, what is this 1, 2 to 7?*

O: *The competition: [races] 1, 2 and 3.*

I: *Now, please explain the ticks [you made] in the table.*

O: *This is, here in the first competition, the winner is Rita; [in] the second competition [the winner] is also Rita; [in] the third [competition the winner] is Rita, [in] the fourth [competition the winner] is Rita, and [in] the fifth [competition the winner] is Maria; [in] the sixth [competition the winner] is Maria, and [in] the seventh [competition the winner] is Maria.*

I: *So, the ticks show the winner; how do you know that the winner in the first competition is Rita?*

O: *I thought these* [pointing to the times] *are scores.*

I: *Scores? Ok, it means you consider 15.3 a score and [think] Rita is the winner because that [number] is the biggest?*

O: *Yes*.

From this interview, Oktav seemed to base her selection on the statistical concept of the mode; whoever had ‘won’ the most races, based on her interpretation of the scoring, should be selected to compete in the next championship. She determined who had won each race based on that thinking and thus thought Rita had won four times, Maria had won three times and Sarah had never won. Without using any further reasoning, Oktav selected Rita because she had won four times.

However, the interviewer did not want to end the interview at this point. The question remained as to whether the context and table were difficult to understand or whether Oktav's mistake was due to a lack of knowledge. Therefore, the interviewer asked Oktav to reread the text and explain what she could comprehend, giving her the opportunity to correct her challenges in data reading by recognising that the numbers in the table were times instead of scores. Oktav corrected herself after realising that the winner in each race should be the one with the lowest time:

I: *OK, now, can you please read the text again and please interpret it?*

O: [Reads the text above the table]

I: *Now I ask again, what is this 1 to 7?*

O: A *competition.*

I: *And the three students are Sarah, Rita and Maria?*

O: *Yes.*

I: *What are they—the numbers in the table?*

O: *Scores.*

I: *So, the score is 15.2?* [pointing to Sarah]

O: *Hmmm, wait a minute, that's wrong.*

I: *What do you mean? What do you understand now?*

O: *This should be times, showing times in seconds. So, [it] must be her* [pointing to Maria] *because her time is the quickest, the shortest.*

I: *Can you please explain again?*

O: *In the text, it is written: 'this table shows times (in seconds)', so these are times. First, I thought they were scores. So, the winner should be Maria [in the first race] because her time is the shortest.*

From this interview, it can be concluded that the challenges Oktav encountered was not caused by an unclear text or table, but by a careless reading error. Therefore, no refinements were needed for this item.

4.3.1.4 Revised Items

Table 4.13 summarises the overall findings from Pilot Interview II about the 12 students' understandings and challenges of The 100-Metre Race item.

Table 4.13

Students' Responses to The 100-Metre Race Item

Response	Year 9	Year 12
Understanding	The shortest total time: *Sarah, because in seven races the time she took is shorter than Rita's or Maria's.*	Mean and trend: *I will select Sarah … because the mean times of Sarah and Maria are [the] same, but Sarah [has] an increase [in speed].* *I select Sarah because Maria took a longer time than in previous races, although her mean is the same as Sarah's.* Trend and the last race: *Sarah has a constant increase [in speed] and Rita otherwise, although they both have the same time in Race 7.*
Item-caused challenges	N/A	N/A
Student-caused challenges	Erroneous calculation: Incorrectly finding the total time for Maria to be 87 seconds instead of 103 seconds. *I will select Maria because she has the quickest mean. Sarah 18.7, Rita 733.6 and Maria 12.8, so Maria.*	The greater the trend, the quicker the runner: *Maria always increases, from Race 1 to 7.* The longest time is the winner: *Rita won races 1 to 4 while Maria won races 5 to 7.* The winner has the highest 'score' in the last three races: *Maria won … the last three races, 5 to 7, so I select her.*

Table 4.13 (continued)

Response	Year 9	Year 12
Student-caused challenges	The winner has the highest ‘score’ or time: *Among Sarah 15.2, Rita 15.3, Maria 14.0, the choice was Rita as she has the highest time.* The winner has the highest mean: *Among Sarah 14.7, Rita 14.9, Maria 14.7, the choice was Rita as she has the highest mean.* The winner is the one who has the highest ‘score’ most often: *Rita, because she won the most (four out of seven races).*	The best has the highest mean: *I choose Rita because she has the highest mean.* Continuous increase: *Maria shows continuous increase in races 5 to 7.*

Based on the findings of the students’ interview and their written test answers to The 100-Metre Race item, it was concluded that this item and its features (text, table and question) are comprehensible for the students. Students’ challenges or incorrect responses were merely caused by their lack of knowledge or careless reading errors. Students’ inability to make sense of the context of a running competition, interpret data in the table, determine the winner (using the shortest time, total time or mean, or the ‘getting-faster’ trend) are a few examples of their lack of knowledge. Therefore, no revision was needed, and this item was used for the SL test.

Eventually, similar procedures were applied to all 14 items. The findings motivated three types of refinements to the items. The first type of change was applied to the provided contextual texts. These changes varied from merely adding a comma (or otherwise changing the punctuation) to rephrasing words and phrases that were ambiguous or difficult to understand, shortening items that were too long and replacing or adding wording or numbers to provide students with clearer context. The second type of amendment was making alterations to the items’ graphical or tabular representations. This amendment aimed to improve the clarity and simplicity of those graphs and tables that might be misinterpreted. Finally, the third type of item refinement was applied to the questions' wording. Extra instructions were added to some items to prompt students to provide long responses. Table 4.14 shows the refinements made to one example item for each skill. The other refined items are available in Section 4.4.

Table 4.14

Assessment Item Refinement after Pilot Interview II

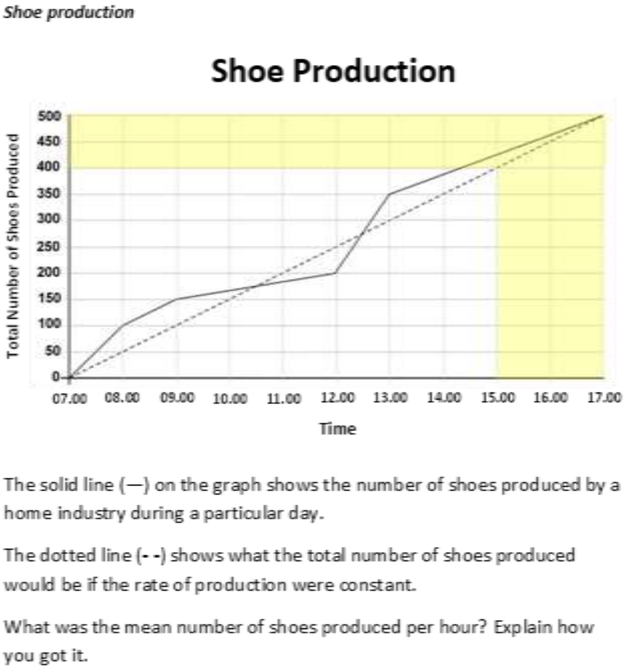
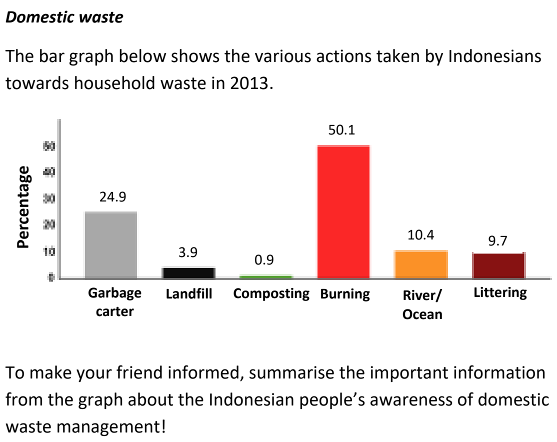
Final refinements	Revised item after Pilot Interview II
The context was retained. The graph was modified so that students could respond at the highest level (critical mathematical). It was expected that changing the range of the x-axis (from 8 to 10 hours) and the maximum of the y-axis (from 400 to 500) would allow students to compare the numbers for 12.00–13.00 and 13.00–17.00 and observe that 150 shoes were produced in both periods. This modification was applied to The Most Production item which shares the same context with The Production Mean item. The question was retained.	Interpreting item (I)  *Shoe production* The solid line (—) on the graph shows the number of shoes produced by a home industry during a particular day. The dotted line (- -) shows what the total number of shoes produced would be if the rate of production were constant. What was the mean number of shoes produced per hour? Explain how you got it.

Table 4.14 (continued)

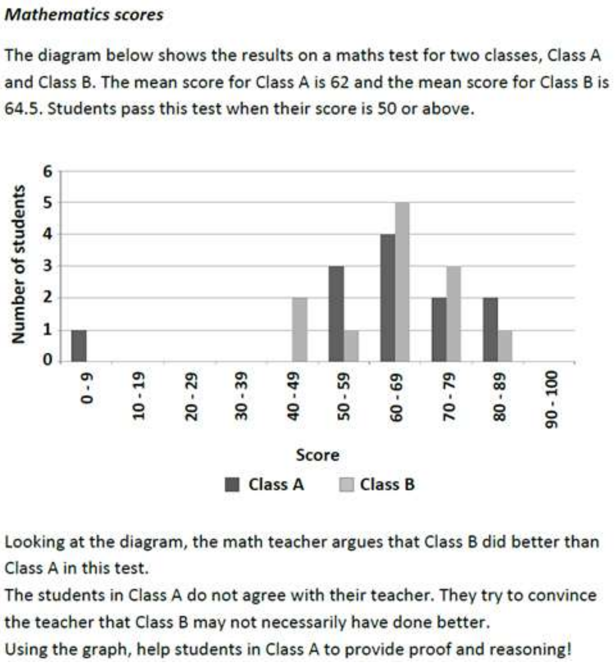
<table>
<tr><th>Final refinements</th><th>Revised item after Pilot Interview II</th></tr>
<tr><td>The context and question were both retained. A slight change was made to the bar colours, but this would not affect students’ understanding of this item.</td><td>Communicating item (C)



Domestic waste

The bar graph below shows the various actions taken by Indonesians towards household waste in 2013.

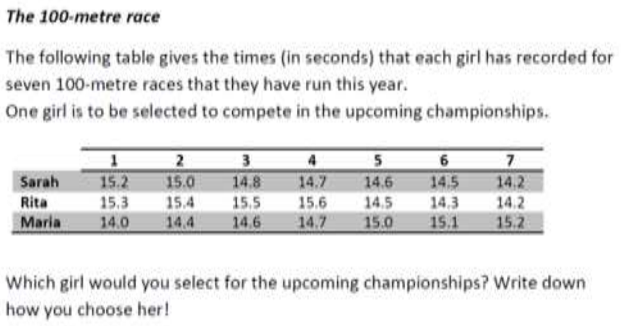
To make your friend informed, summarise the important information from the graph about the Indonesian people’s awareness of domestic waste management!</td></tr>
<tr><td>The context and question were both retained. A slight change was made to the bar colours, but this would not affect students’ understanding of this item.</td><td>Evaluating item (E)



Mathematics scores

The diagram below shows the results on a maths test for two classes, Class A and Class B. The mean score for Class A is 62 and the mean score for Class B is 64.5. Students pass this test when their score is 50 or above.

Looking at the diagram, the math teacher argues that Class B did better than Class A in this test.
The students in Class A do not agree with their teacher. They try to convince the teacher that Class B may not necessarily have done better.
Using the graph, help students in Class A to provide proof and reasoning!</td></tr>
<tr><td>The context, graph and question were all retained.</td><td>Decision-making item (D)



The 100-metre race

The following table gives the times (in seconds) that each girl has recorded for seven 100-metre races that they have run this year.
One girl is to be selected to compete in the upcoming championships.

<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr>
<tr><td>Sarah</td><td>15.2</td><td>15.0</td><td>14.8</td><td>14.7</td><td>14.6</td><td>14.5</td><td>14.2</td></tr>
<tr><td>Rita</td><td>15.3</td><td>15.4</td><td>15.5</td><td>15.6</td><td>14.5</td><td>14.3</td><td>14.2</td></tr>
<tr><td>Maria</td><td>14.0</td><td>14.4</td><td>14.6</td><td>14.7</td><td>15.0</td><td>15.1</td><td>15.2</td></tr>
</table>

Which girl would you select for the upcoming championships? Write down how you choose her!</td></tr>
</table>

4.3.2 Results of Piloting and Refining Cognitive Interview Protocol

The initial version of the interview protocol was piloted (Pilot Interview I), revised, and re-piloted (Pilot Interview II) before being employed for data collection. The two pilot interviews had different aims for improving the interview protocol. Pilot Interview I was conducted to validate both the interview protocol and the assessment items, whereas Pilot Interview II was a small-scale interview to validate the interview protocol based on the students' written responses. These pilot interviews improved the process of conducting the actual interviews with the most effective wording of interview questions. The pilot interviews showed that using semi-formal language during an interview has benefits, such as engaging students to express their thoughts freely. In other words, no sociocultural barriers existed between the interviewer and students, making the interview flow naturally, like two people sharing instead of interviewing and being interviewed. More importantly, it helped the researcher attain the interview goal: the participants were able to verbalise the mental activities they experienced when working on a test (Conrad & Blair, 1996), expressing their thought processes and verbal reasoning (Desimone & Le Floch, 2004).

In addition to guiding the item refinement (see Section 4.3.1), piloting informed the logistics of conducting the actual interview: the time allocation, interview process and script. The pilot interviews showed that total time allocated to interviewing each student could be predicted better. The piloting informed the researcher that interviewing each item took about five to seven minutes. Thus, it took, in total, around an hour and a half to interview each student. Fourteen items were included in the interview, and piloting indicated that the students became bored. Consequently, the questions were revised to avoid utilising the four problem-solving processes (comprehending the item, retrieving relevant information, making a judgement based upon the recall of knowledge and mapping the answer onto the reporting system) strictly. Instead, once the students were familiar with the interview steps, they were just asked to explain their written answers, and the interviewer responded with probing questions. This flexibility was not anticipated to be a problem, because the interviewer would be adapting the interview protocol to each student based on the conjectures being generated about the student's cognitive processes for each item. Moreover, the aim was to reveal the students' thought processes. Table 4.15 exemplifies the revised interview protocol for The 100-Metre Race item.

Table 4.15

Revised Interview Script

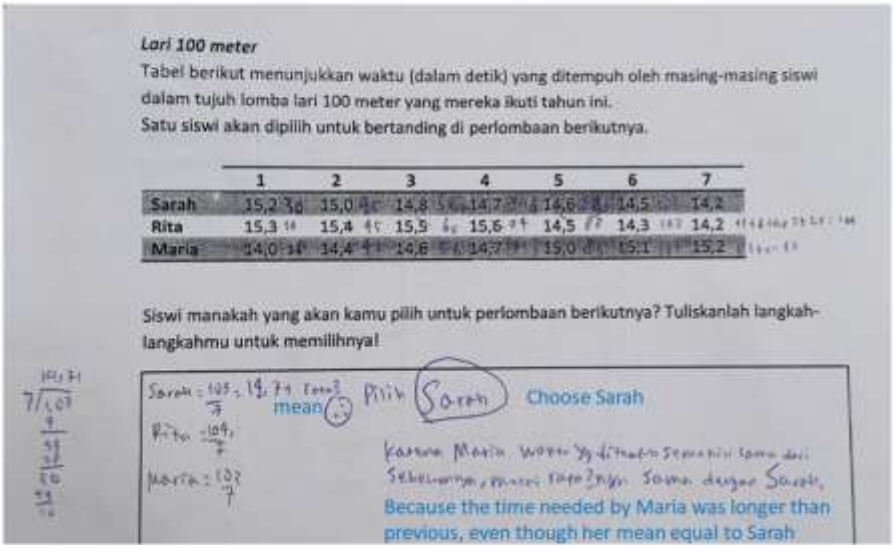
<table>
<tr><th>Step</th><th>Script</th></tr>
<tr><td>Beginning the interview
(4 minutes)</td><td><i>Before we begin this interview, let's have a proper introduction. My name is Badrun, and you are</i> [insert interviewee's name]. <i>It's nice to meet you.</i>
<i>First of all, thanks so much for participating in this interview. Your input is going to help me understand what's happening in your mind when you're working on data-based problems. It's not about right or wrong answers, so there's no need to be scared. What you say during this interview will be recorded. Is that OK</i> [insert interviewee's name]<i>? Thank you.</i>
<i>During this interview, please think aloud as you're solving the problems. That means say anything—whatever you think; I'm interested in hearing all of your thoughts and reactions.</i> <b>(Repeat and emphasise this information.)</b>
<i>We will now begin this cognitive interview at time</i> [insert time].</td></tr>
<tr><td>Conducting the interview
(5–7 minutes per item)</td><td>
<b><i>Lari 100 meter</i></b>
Tabel berikut menunjukkan waktu (dalam detik) yang ditempuh oleh masing-masing siswi dalam tujuh lomba lari 100 meter yang mereka ikuti tahun ini.
Satu siswi akan dipilih untuk bertanding di perlombaan berikutnya.
<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr>
<tr><td>Sarah</td><td>15,2</td><td>15,0</td><td>14,8</td><td>14,7</td><td>14,6</td><td>14,5</td><td>14,2</td></tr>
<tr><td>Rita</td><td>15,3</td><td>15,4</td><td>15,5</td><td>15,6</td><td>14,5</td><td>14,3</td><td>14,2</td></tr>
<tr><td>Maria</td><td>14,0</td><td>14,4</td><td>14,6</td><td>14,7</td><td>15,0</td><td>15,1</td><td>15,2</td></tr>
</table>
Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!
Sarah = 103/7 : 14,71 (rata2) mean Pilih Sarah Choose Sarah
Rita = 104,/7
Maria = 103/7
Karena Maria waktu yg ditempuh semakin lama dari sebelumnya, meski rata2nya sama dengan Sarah,
Because the time needed by Maria was longer than previous, even though her mean equal to Sarah
<i>Can you tell me what you can understand from The 100-Metre Race item? Please say it aloud!</i></td></tr>
</table>

Table 4.15 (continued)

Step	Script
Conducting the interview (5–7 minutes per item)	**Probe:** *Did you understand all the information above the table and in the table?* **If no, probe:** *Which particular information in this item was difficult to understand?* **If yes, probe:** *Can you give me a short explanation?* *That's great. Thinking out loud like this is just what I need.* *Next, what is this question asking?* **Probe:** *Okay, can you explain the way you chose the best runner?* *That's great.* **Probe:** *Can you go over that again? What was your answer? Why did you choose her?* *The way you calculated looks interesting. Can you please explain it to me?* *Well done.* (This series of questions can be further developed based on the students' responses during the interview.)
Closing the interview (1 minute)	*Thank you for taking the time to participate in this interview. If you have any comments to share, please feel free.* *That's the end of this interview, and I will now stop the recording at time* [insert time].

4.3.3 Results of Piloting and Refining Descriptors

All the findings from the three-stage piloting also contributed to the development of the component-based item descriptors. Both the pilot interviews and test greatly contributed to improving on the initial component-based descriptors that had previously been theoretically developed (Chapter 2). In particular, the students' written and spoken responses from the piloting provided substantial evidence. To exemplify the refinement process, Table 4.16 summarises the written and spoken responses of three students (Amal, Zainuddin and Oktav), which were illustrated in the item refinement section (Section 4.3.1).

Table 4.16

Students' Responses as Empirical Evidence for Item Descriptor Refinement

Student	Component	Level
Amal	*Text and context*: He understands that three runners are competing in seven races, and one best runner has to be chosen; he knows that in running competitions, the winner is the one with the shortest time.	Critical mathematical
	Representation: He can see that the table consists of times taken by three runners to reach the finish line in seven races; he can relate these numbers to the context, in which the shorter the time, the quicker the runner reached the finish line.	Critical mathematical
	Statistical-mathematical knowledge: He applies mean and trend as selection methods; he shows sophisticated number sense while calculating the mean.	Critical mathematical
Zainuddin	*Text and context*: He understands that three runners are competing in seven races and one best runner has to be chosen; however, he incorrectly relates the winner in a running competition to other competitions in which the highest number is the winner.	Inconsistent
	Representation: He can see that the table consists of times taken by three runners to reach the finish line in seven races. However, his interpretation is incorrect, influenced by his incorrect contextual understanding (he thinks the longer the time, the quicker the runner reached the finish line).	Inconsistent
	Statistical-mathematical knowledge: He applies the mean as a selection method and performs the correct calculation for the total time and the mean.	Consistent non-critical
Oktav	*Text and context*: She seems to understand that there are seven races and three runners competing; however, she fails to understand that in a running competition, the numbers recorded are not scores.	Informal
	Representation: She misinterprets the data in the table as the runners' scores, leading her to find the one with the highest score.	Informal
	Statistical-mathematical knowledge: She uses the mode to choose the best runner, but her inappropriate contextual understanding of the winner results in the wrong choice.	Between informal and Inconsistent

As observed from Table 4.16, responses from students could be used to enrich the previous descriptors. For example, The 100-Metre Race item descriptor was modified in this way to produce the final descriptor that would serve as a scoring guide for the actual data collection (see Table 4.17). The other items underwent the same refinement process.

Table 4.17

Development of Component-based Item Descriptors for The 100-Metre Race Item

Level	Initial component-based item descriptor	Revised component-based item descriptor
Idiosyncratic	*Text and context*: Students interpret the context of running using personal experience and use it to choose one of the runners. *Representation*: Students choose inappropriate data from the table. *Statistical-mathematical knowledge:* Students make calculations based on the numbers in the table but not relating them to the question.	*Text and context*: Students involve personal experience in understanding the context of a running competition and fail to link the information in the text with the numbers in the table. *Representation*: Students choose inappropriate data from the table; for example, they use 1–7 (i.e., Competitions 1–7) instead of the times in those even races. *Statistical-mathematical knowledge*: Students make no calculations; or if they do, the calculations do not relate to the question or cannot be understood. They might choose one runner, but without providing statistical-mathematical justification.
Informal	*Text and context*: Students interpret the context of running using everyday experience (the winner has the biggest number) and use it to choose one of the runners. *Representation*: Students use the data for three runners in seven races, but do not interpret the values as times. *Statistical-mathematical knowledge:* Students use the total of the data to choose the best runner (the one with the longest total time).	*Text and context*: Students analogise the context of choosing one runner from three with other contexts in which the best is the biggest or the highest. *Representation*: Students misinterpret the time in the table as something else, such as a score, so they focus on the highest number for each race or the total for each runner. *Statistical-mathematical knowledge*: Students still show no knowledge of statistical-mathematical concepts in choosing one of the runners; errors include choosing the biggest data.

Table 4.17 (continued)

Level	Initial component-based item descriptor	Revised component-based item descriptor
Inconsistent	*Text and context*: Students understand partially the context of a running competition but still use an informal interpretation. *Representation*: Students read the data in a table that displays the times for three runners over seven races but fail to recognise the relationships between the data. *Statistical-mathematical knowledge:* Students use the mean or mode to figure out who is the best runner but inappropriately choose the largest mean or the longest time.	*Text and context*: Students start by comprehending that three runners are competing in seven races, and one best runner has to be chosen; however, they still relate the winner in a running competition to other competitions in which the highest number is the winner. *Representation*: Students can interpret the data in the table as presenting the times for the three runners over seven races but fail to identify the relationships between the data; errors include focusing on the runner in each race with the biggest number. *Statistical-mathematical knowledge*: Students start by using statistical concepts, such as mean or mode, to choose one runner, but inappropriately choose the runner with the highest mean or find the highest values for each race to choose the runner who ‘won’ the most races. The calculations they perform show a significant amount of error.
Consistent non-critical	*Text and context*: Students appropriately understand the context of a running competition in which the winner is the one with the shortest time. *Representation*: Students read the data in a table that displays the times of three runners in seven races and recognise the relationship between them. *Statistical-mathematical knowledge:* Students use the mean to choose two runners with the same average.	*Text and context*: Students understand that three runners are competing in seven races, and one best runner has to be chosen. They also understand the context of a running competition in which the winner is the one with the shortest time in each race or the lowest total time or mean. *Representation*: Students can interpret the data in the table as presenting the times of three runners in seven races and identify the relationship between them by focusing on the lowest number in each race or the total for each runner. *Statistical-mathematical knowledge*: Students choose one runner using a

Table 4.17 (continued)

Level	Initial component-based item descriptor	Revised component-based item descriptor
		correct statistical concept such as mean, mode, trend or total time, performing the correct calculation.
Critical	*Text and context*: Students critically understand the context of a running competition. *Representation*: Students recognise that the best runner is the one with the lowest mean. *Statistical-mathematical knowledge:* Students choose one of two runners with the same average and justify that choice by, for example, choosing the one who won most often.	*Text and context*: Students demonstrate critical contextual understanding when they discover two runners with the same mean, but only choose one of them. *Representation*: Students can critically relate the context to the data in the table by focusing on the shortest time, total time or trend of each runner. *Statistical-mathematical knowledge*: Students show critical thinking when they find two runners with the same mean and use additional justification to choose one.
Critical mathematical	*Text and context*: Students understand critically the context of running competitions and factors to be considered in choosing the best runner. *Representation*: Students summarise the data in the table (such as mean, variation and trend) to select the best runner. *Statistical-mathematical knowledge:* Students choose one out of two runners with the same average and justify the choice, for example, by comparing the trend.	*Text and context*: Students demonstrate critical contextual understanding that relates to their critical understanding of representation and statistical-mathematical knowledge. *Representation*: Students critically interpret the data in the table and are able to see that the best runner can be selected in several ways, such as by mean, trend or variation. *Statistical-mathematical knowledge*: Students show critical thinking when they find two runners with the same mean, by using the mode, trend or distribution as an alternative selection method. Students also show sophisticated number sense and accurate calculating.

4.4 Test Items

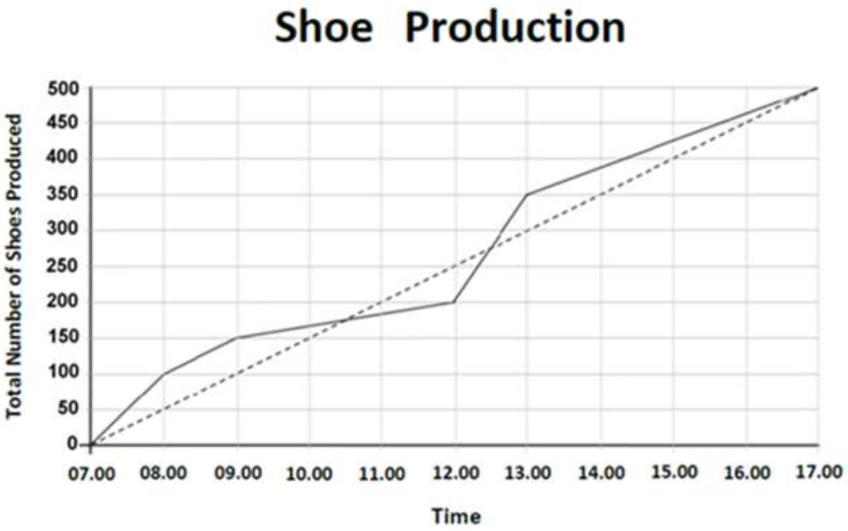
The three stages of piloting resulted in 10 assessment items for the actual data collection. Four of the initial items were removed because the findings from the pilot proved that they could not be improved to cover all six levels. The 10 selected items comprised four interpreting items (The Production Mean, The Most Production, In Which Month? and How Many Students?), two communicating items (Domestic Waste and YouTube Viewers), two evaluating items (Mathematics Scores and The Employees) and two decision-making items (The 100-Metre Race and Which Motorcycle?). Eight of these ten items are illustrated below and see Appendix F for the other two interpreting items (In Which Month? and How Many Students?).

4.4.1 Items of Interpreting and Communicating Skills and Their Three Components

There were four items designed to assess students' interpreting skills. Two of these four items share the same shoe production context (Figure 4.3). The first item (The Production Mean) asked students, '*What was the mean number of shoes produced per hour? Explain how you got it!*' It was intended to assess students' comprehension of the mean as a measure of central tendency obtained from a line graph. The second item (The Most Production) asked the students, '*During which time interval were the most shoes produced? Explain and relate your answer to the shoe production graph!*' It aimed to assess the students' ability to locate the interval with the highest values in a line graph. For these items, students were required to show contextual, graphical and statistical-mathematical understanding. The students were expected to be able to justify that the mean production is the per hour increase shown by the dotted line. Further, the students were expected to reason that the most production occurred between the two data points connected by the line with the sharpest slope.

Figure 4.3

Context and Graph for the Two Interpreting Skill Items



The solid line (——) on graph shows the number of shoes produced by a home industry during a particular day.

The dotted line (- - - -) shows what the total number of shoes produced would be if the rate of production were constant.

Students must understand the three SL components to solve these two interpreting items. Table 4.18 summarises the three SL components students need to comprehend to interpret the information in the abovementioned problems. In terms of text and context, the students should understand that two lines represent the same data for different purposes. Further, there is more textual information that students need to consider, such as ‘constant’ increase and mean per hour. Students must apply their contextual understanding when making sense of the data presented in the line graph. For example, they must realise that the two lines represent the same discrete data shown cumulatively. Finally, the students should demonstrate an understanding of the mean concept and procedure and locate the period during which the most production occurred.

Table 4.18

Assessment of the Three Components of the Interpreting Skill

	Text and context	Representation	Statistical-mathematical knowledge
The Production Mean	Two lines represent the same data: the solid line represents the raw data, and the dotted line represents the processed data. The word ‘constant’ and the idea of increase.	Line graph conventions (such as the labels of the x and y axes). The data is discrete and cumulative. Data points, data increments and constant increase.	The concept of mean obtained from either a formula or a graph. Number operations, especially addition and division.
The Most Production	Two lines represent the same data: the solid line represents the raw data, and the dotted line represents the processed data. The word ‘constant’, the idea of increase, time interval, most production.	Line graph conventions (such as the labels of the x and y axes). The data is discrete and cumulative. Data points, data increments, constant increase, and slope.	Proportional comparison to determine the most production within the shortest time. The statistical meaning of the sharpest slope.

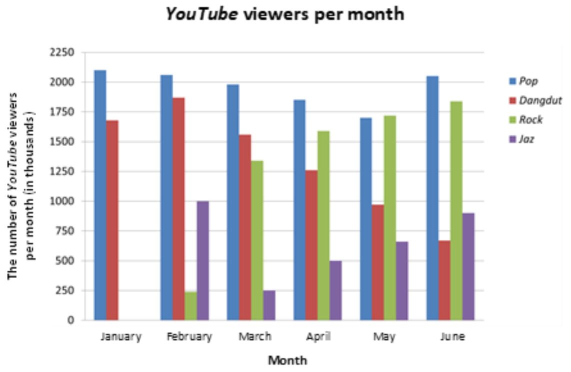
Regarding communicating, two items were designed to assess this skill (see Figure 4.4). These two items are YouTube Viewers and Domestic Waste, both of which involve data in bar graphs. For YouTube Viewers, the students were expected to provide summary information, showing comparisons, relationships, trends and the most relevant data, from a graph showing the number of YouTube viewers of four bands throughout six months. Similarly, for the Domestic Waste item, the students were expected to summarise the six actions Indonesians took towards their domestic waste. To do so, they needed to critically summarise important information by grouping, comparing, contrasting, including the most relevant data and making summary statements regarding people’s awareness of managing domestic waste.

Figure 4.4

Two Communicating Skill Items

YouTube viewers

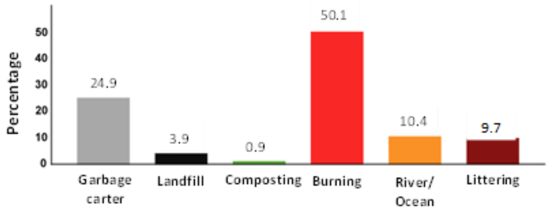
In January, new singles by the bands *Pop* and *Dangdut* were released. In February, singles by the bands *Rock* and *Jazz* followed. The following graph shows these bands’ number of *YouTube* viewers from January to June.



Please write a summary of the important information that you understand from the graph!

Domestic waste

The bar graph below shows the various actions taken by Indonesians towards household waste in 2013.



To make your friend informed, summarise the important information from the graph about the Indonesian people’s awareness of domestic waste management!

The three-component understandings that students are expected to need to solve the two communicating items successfully are presented in Table 4.19. In terms of text and context, the students should understand that the four bands released their singles in different months (for YouTube Viewers) and that among the six actions towards household waste, some are proper, and some are improper (for Domestic Waste). Such understanding is important to make sense of the data in bar graphs. In addition, knowledge of statistical ideas, such as trends and means, is also important when interpreting the details of bar graphs.

Table 4.19

Assessment of the Three Components of the Communicating Skill

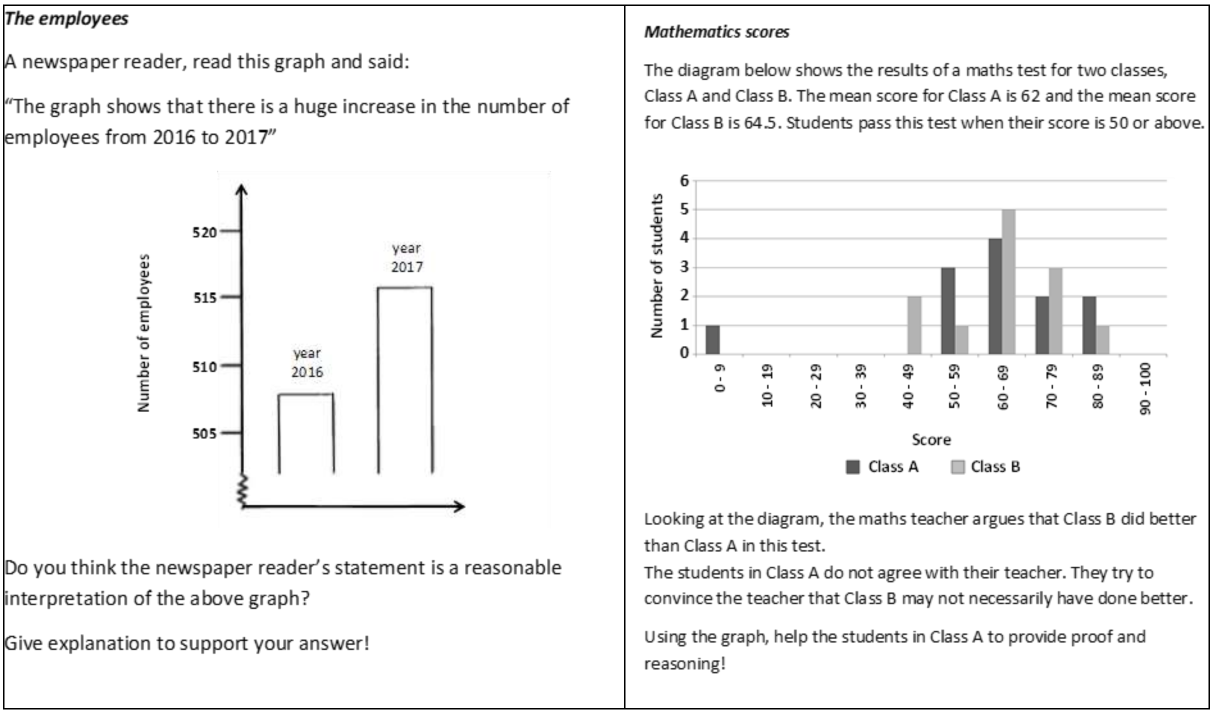
	Text and context	Representation	Statistical-mathematical knowledge
YouTube Viewers	Two bands released singles in January, and two bands did so in February. Understand the instruction to write summary information.	The label of the y-axis shows the number of YouTube viewers in thousands. The legend shows the four bands in different colours. Only two bands released their singles in January.	Identify the overall trend for each band. Comparing the most and least watched. Using average in a summary. Grouping the bands with the same pattern or trend.
Domestic Waste	The data was collected in 2013. There were six actions; some are proper, and some are improper. Understand the instruction to write summary information on people's awareness.	The label of the y-axis shows the percentage for each action. The number above the bars is the exact percentage	Comparing the most and least common action Grouping and comparing the actions into two categories (proper and improper).

4.4.2 Items of Evaluating and Decision-Making Skills and Their Three Components

There were two items selected to assess the students' skills in evaluating data-based claims or arguments (see Figure 4.5). These items are The Employees and Mathematics Scores, both of which involve data in bar graphs. For The Employees item, students are expected to challenge the misleading claim made by a newspaper reader. A newspaper reader has claimed that a 'huge increase' in the number of employees from 2016 to 2017 is shown in the bar graph. For the Mathematics Scores item, the students are expected to challenge a misleading claim made by a mathematics teacher. The teacher claims that the students in class B performed better than the students in class A on a mathematics test. These two problems are sufficiently open ended to enable students to refute the claims with various pieces of evidence.

Figure 4.5

Two Evaluating Skill Items



The employees

A newspaper reader, read this graph and said:

"The graph shows that there is a huge increase in the number of employees from 2016 to 2017"

Number of employees
520
515
510
505
year
2016
year
2017

Do you think the newspaper reader's statement is a reasonable interpretation of the above graph?

Give explanation to support your answer!

Mathematics scores

The diagram below shows the results of a maths test for two classes, Class A and Class B. The mean score for Class A is 62 and the mean score for Class B is 64.5. Students pass this test when their score is 50 or above.

Looking at the diagram, the maths teacher argues that Class B did better than Class A in this test.
The students in Class A do not agree with their teacher. They try to convince the teacher that Class B may not necessarily have done better.

Using the graph, help the students in Class A to provide proof and reasoning!

Students must understand the three SL components to solve these two evaluating items successfully. Table 4.20 summarises the three-component comprehension students need to answer the two items. In terms of text and context, the students should understand that they were asked to challenge the existing claim made by the newspaper reader (for The Employees item) and the claim made by the mathematics teacher (for the Mathematics Scores item). Further, more information from the text must be considered, such as the phrase 'huge increase' (The Employees) and the passing grade and the mean for the two classes (Mathematics Scores). Students need to apply their contextual understanding when making sense of the data presented in the bar graphs. For example, they need to consider that the *y-axis* in The Employees item has a discontinuity while the *x-axis* labels in the Mathematics Scores item show interval data. Finally, the evidence they use to challenge the claims must consist of statistical and mathematical ideas, such as benchmark and percentage for The Employees item and mean and outlier for the Mathematics Scores item.

Table 4.20

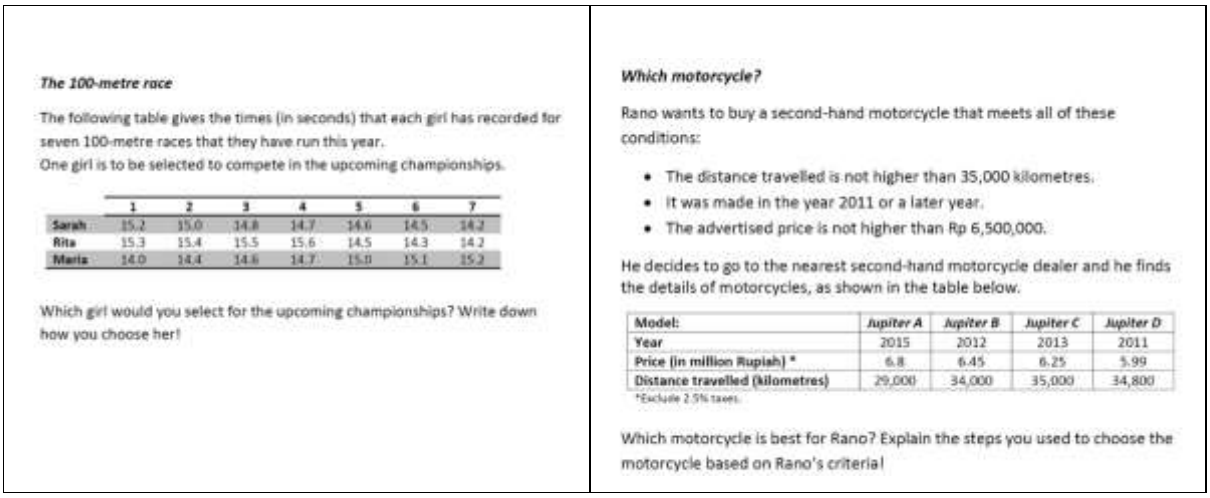
Assessment of the Three Components of the Evaluating Skill

Component	Text and context	Representation	Statistical-mathematical knowledge
The Employees	The number of employees in each of two years is compared. The real increase versus 'huge increase' in the number of employees. There is a newspaper reader's claim to be challenged.	The bar graph convention (such as the labels for the x and y axes). The discontinuity in the y-axis.	Number operation (particularly addition and subtraction). Predicting the number of employees. Benchmark (such as percentage and the need for more data) to determine whether the increase is 'huge'.
Mathematics Scores	Two classes are compared. The information provided in the text for the two classes (mean and passing grade). There is a claim made by a mathematics teacher that must be challenged.	Bar graph conventions (such as the labels of the x and y axes). The x-axis shows interval data.	Mean Minimum passing grade Outlier

Finally, two items were used to assess the students' skill at making decisions based on data (see Figure 4.6). These items are The 100-Metre Race and Which Motorcycle?, both of which involve data in a table. In The 100-Metre Race item, the students were expected to be able to select the best runner among three runners to compete in the upcoming championship. As evidence to support their selected runner, they could use the data presented in the table. The second item, Which Motorcycle?, was developed to assess the students' capacity to make decisions based on three numerical conditions. The students were expected to be able to select one of four motorcycles from a table based on three criteria that are set in a numerical context. They also needed to consider that a tax was not included in the prices. These two problems are sufficiently open ended for students to make well-informed decisions based on various pieces of evidence.

Figure 4.6

Two Decision-Making Skill Items



The 100-metre race

The following table gives the times (in seconds) that each girl has recorded for seven 100-metre races that they have run this year.
One girl is to be selected to compete in the upcoming championships.

	1	2	3	4	5	6	7
Sarah	15.2	15.0	14.8	14.7	14.6	14.5	14.2
Rita	15.3	15.4	15.5	15.6	14.5	14.3	14.2
Maria	14.0	14.4	14.6	14.7	15.0	15.1	15.2

Which girl would you select for the upcoming championships? Write down how you choose her!

Which motorcycle?

Rano wants to buy a second-hand motorcycle that meets all of these conditions:

- The distance travelled is not higher than 35,000 kilometres.
- It was made in the year 2011 or a later year.
- The advertised price is not higher than Rp 6,500,000.

He decides to go to the nearest second-hand motorcycle dealer and he finds the details of motorcycles, as shown in the table below.

Model:	*Jupiter A*	*Jupiter B*	*Jupiter C*	*Jupiter D*
Year	2015	2012	2013	2011
Price (in million Rupiah) *	6.8	6.45	6.25	5.99
Distance travelled (kilometres)	29,000	34,000	35,000	34,800

*Exclude 2.5% taxes.

Which motorcycle is best for Rano? Explain the steps you used to choose the motorcycle based on Rano's criteria!

Students must understand the three SL components to solve the two decision-making items successfully. Table 4.21 summarises the three components that students must comprehend to make informed decisions. In terms of text and context, the students must understand that they are being asked to choose one best runner (The 100-Metre Race) and the best motorcycle (Which Motorcycle?). Further, there is more contextual information that students need to consider, such as that the shorter the time, the better the runner (The 100-Metre Race), and the three numerical conditions that need to be met (Which Motorcycle?). Students need to apply their contextual understanding when making sense of the data presented in the tables. For example, they need to consider that the table for the three runners presents the runners' times, while the table in Which Motorcycle? presents numerical data about the four motorcycles. Finally, their selection of a runner or motorcycle must be supported by relevant reasoning using statistical and mathematical ideas, such as measures of centre and trend for The 100-Metre Race item and percentage and inequality for Which Motorcycle?

Table 4.21

Assessment of the Three Components of the Decision-Making Skill

Component	Text and context	Representation	Statistical-mathematical knowledge
The 100-Metre Race	Three runners are compared. Choosing one runner. A smaller number means a faster run.	Table conventions (rows and columns) and the data presented (time in seconds).	Mean, mode, trend and number sense.
Which Motorcycle?	The three numerical conditions Four motorcycles are compared. Additional tax Choosing the best motorcycle	Table conventions (rows and columns) and the data presented. The price is in millions.	The concept of inequality Percentage Comparison

4.5 Chapter Summary

This chapter has described the details of the instrument development and piloting processes. These details were intended to provide evidence of the instruments' validity. Three stages of initial development and three stages of piloting helped to produce valid instruments: the items for an SL test, component-based item descriptors for scoring guide and the interview protocol for the follow-up interview. These items were not designed to assess one of the six hierarchical levels. Instead, component-based item descriptors were devised to capture responses spanning all six hierarchical levels. Thus, no further quantitative validation was needed to ensure if the items were normally distributed across the hierarchy. As the three instruments were developed following the proposed assessment framework, this chapter has additionally demonstrated that such assessment framework can be used to assess students' SL. The next chapter describes how these instruments, and the framework are used to collect and analyse data.

Chapter 5: Methodology

This chapter details the methodology employed in this study. Section 5.1 presents the theories that underpin the study's quantitative and qualitative cross-sectional design. Section 5.2 presents the overall sequence of this study. Section 5.3 contains information about the participants from both the quantitative and qualitative studies. Section 5.4 provides an explanation of the data collected from participants and the data collection methods. Section 5.5 describes the data analysis techniques. Finally, Section 5.6 provides a chapter summary.

5.1 Theoretical Underpinning of Cross-Sectional Design

As summarised in previous chapters, the aim of this study was to investigate Indonesian high school students' SL. The first two research questions of the study are quantitative, aimed at discovering the students' SL levels and the differences between students' SL. Specifically, the research questions seek to understand what SL levels Indonesian high school students are capable of attaining and whether the students' SL levels develop with age, differ by gender or differ according to the students' backgrounds (school type, school status and city of origin). The two remaining research questions of the study are qualitative and seek to establish what challenges students encounter and what understandings they have of three SL components when responding to data-based items. Based on the aim and research questions, a cross-sectional study design was deemed the most appropriate.

A cross-sectional study is a type of observational study that examines variables at a single point in time (Levin, 2006; Mann, 2003). In a cross-sectional study, the researcher can provide a snapshot of what is happening in a specific group of people (Bourque, 2003) or across multiple groups simultaneously (Montague & Van Garderen, 2003). From those groups, cross-sectional data are obtained (Spector, 2003). Using such data, there are at least two feasible analyses: assessing the strength of associations between observed variables and testing the significance of group differences (Whalley, 2006). The difference between two groups of people may follow one of three patterns: a significant increase, a significant decrease or no significant change (Sutton, 2000). A cross-sectional study was deemed appropriate for the current study because 1) the participants of this study had heterogeneous characteristics (Section 5.3), 2) the data were collected in a short time frame (Section 5.4) and 3) the differences between groups were investigated (Sections 5.5.1.3–5.5.1.4).

Although a quantitative cross-sectional study may be part of an observational study (Mann, 2003), other versions have commonly been used in studies. When ‘cross-sectional studies’ was searched for on research databases such as Web of Science and Google Scholar, the findings suggested that this study design frequently appears in health-related studies (e.g., Ekanayake et al., 2012; Kesmodel, 2018; Raynes-Greenow et al., 2013; Setia, 2016; Taniyama et al., 2012; Wang & Cheng, 2020). Further searches revealed that there are three types of cross-sectional studies: quantitative, qualitative and a combination of the two. Zheng’s (2015) analysis of the combined type yielded several noteworthy classifications. He reviewed the methodology of 78 articles in health science studies to provide insights into how the combined type of cross-sectional study was conducted in health science. His findings were organised under two major classifications: 1) research design and 2) overall sequences. He identified three types of research design: convergent design, explanatory sequential design and exploratory sequential design. This classification followed Creswell’s (2014) classification of mixed-methods design. In other words, the three research designs correspond to the identified overarching sequences: quantitative and qualitative, quantitative then qualitative, and qualitative then quantitative. Based on these classifications, the current study employed an explanatory sequential design, starting with a quantitative component followed by a qualitative one.

Reviewing the literature revealed that all three types of cross-sectional studies have been employed in education studies, particularly in mathematics and statistics. However, these studies did not specify cross-sectional designs explicitly in their methodologies. For example, Jurdak et al. (2014) provided both quantitative and qualitative results describing Year 4–11 students' reasoning development for pattern generalisation tasks; Ludwig and Xu (2010) presented quantitative results of modelling competencies by grade level, gender and country; and Whitacre et al. (2017) presented students' justifications and methods of reasoning on integer comparisons by grade level. In statistics education, some studies applied various types of cross-sectional analysis. For instance, Mooney (2002) interviewed Years 6–8 students to establish and verify a framework for assessing middle school students' statistical thinking levels, while Aoyama and Stephens (2003) conducted a test and interviews with Years 5–8 students to determine their SL levels. Further, Yolcu (2014) tested students from Years 6–8 to determine the effects of grade level and gender on students' SL test scores.

According to Yorke and Zaitseva (2013), there are at least three significant advantages of employing a cross-sectional design. First, it may be used to rapidly investigate participants from various backgrounds. In this study, the participants were recruited from two distinct grade levels (Years 9 and 12), allowing for investigation of the grade levels' impact on high school students' SL without requiring three years of observation. Second, a cross-sectional design can serve as an adequate substitute for a longitudinal study. The students' SL levels from different grade levels—found by this study—can provide some information about students' performances across those two grade levels. Third, the study's findings may contribute to enhancement-oriented activities. This is important because teachers need to design activities to develop students' SL.

5.2 Study Sequence

To address the research questions of the current study, an explanatory sequential design was used, incorporating a quantitative–qualitative cross-sectional study. The study began with a quantitative component, which investigated the students' SL levels and the differences between students' SL. This was followed by a qualitative component that explored the students' SL-related challenges and understandings to explain the quantitative findings. Zheng (2015) illustrated how a sequence can be used in an explanatory study by referencing Hasan et al. (2014). In their studies, the sequence is 1) quantitative data collection, 2) quantitative data analysis, 3) identification of the need for further exploration, 4) qualitative data collection, 5) qualitative data analysis and 6) integration and interpretation of the quantitative and qualitative results.

That sequence provided the foundation for the current study, in which the quantitative component, in the form of a statistics test, was conducted first, followed by a qualitative interview with some tested students about the assessment items in the test. The test provided quantitative evidence of the students' SL levels, while both the test and the interview provided qualitative information about the students' challenges and understandings when responding to contextual information containing statistics. The overall sequence for this study was 1) quantitative data collection, 2) identification of the need for additional explanation, 3) qualitative data collection, 4) quantitative data analysis and interpretation of quantitative results and 5) qualitative data analysis and interpretation of qualitative results.

5.3 Study Participants

In this study, the participants were selected from Years 9 and 12 rather than from all the grade levels of high school. Although it would have been more comprehensive to include students from a broader range of high school levels (e.g., Years 7–12; Years 7, 9 and 11; or Years 8, 10 and 12), the decision to involve students from only two grade levels was based on three factors. First, Years 9 and 12 mark a crucial age range in the Indonesian school system. Year 9 is the final year of junior high school, and Year 12 marks the end of senior high school and the end of schooling. The details of the Indonesian school system were presented in Chapter 3. Second, Indonesian Year 9 students performed poorly on PISA problems assessing uncertainty and data content over the last two decades (OECD, 2004, 2014). As the PISA result was limited to capturing the national averages (Lowrie & Diezmann, 2009), the involvement of Year 9 students helps to complement the more general reports of PISA.

Third, the SL of Indonesian Year 12 students has rarely been studied; most studies have focused on other high school grades instead (e.g., Fakhmi et al., 2021; Hafiyusholeh, 2015; Hafiyusholeh et al., 2018; Irwandi et al., 2022; Mulya et al., 2018; Oktiviani, 2021; Priyambodo & Maryati, 2019). After graduating from high school, students are unlikely to learn statistics unless they attend a university with a statistics concentration or enrol in certain statistics courses. Nonetheless, they are expected to become statistical citizens, which requires the ability to evaluate statistical information critically (Budgett & Rose, 2017). Consequently, the SL of final-year students would be the best predictor of future statistically literate citizens (Gal, 2002). Identifying the SL level attained by students in Year 12 will provide a snapshot of their SL and their development across a substantial age range of schooling.

Participants were selected by stratified, purposive and convenience sampling. Stratified, purposive and convenience sampling are three common sampling methods identified in quantitative–qualitative cross-sectional studies (Zheng, 2015). Stratified samples are samples within samples (Onwuegbuzie & Collins, 2007), meaning some stratified divisions of the homogeneous sub-group are made within the selected sample (Robinson, 2014; Suri, 2011). The stratification in this study covered the city, school type, school status and gender. The study participants originated from two cities in the province of East Java:

Surabaya (East Java’s capital), which is a representative metropolitan city, and Jombang (the researcher’s hometown), which is a representative non-metropolitan city. In both cities, schools from two different school types were selected: *sekolah* (schools under the auspices of MoEC-RT) and *madrasah* (schools under the auspices of MoRA). The schools also represented two different school statuses: public and private. Based on those stratifications, a total of 16 schools were purposefully chosen. Their accessibility in terms of location and facilities was another consideration in the selection process.

Subsequently, the selection process reviewed the mathematics performance of students in East Java compared to the other 33 provinces and special districts around Indonesia. The students’ performances were compared using the national report from the UN 2019 test (Pusat Penilaian Pendidikan Kemdikbud, 2023). Based on the average score of each province in the mathematics test, the data showed that East Java was in the sixth rank for Year 9 students and in the fifth rank for Year 12 students (see Appendix G). The average score for students in East Java was 48.03 (Year 9) and 41.92 (Year 12 Science majors). Year 12 students in Science majors were selected instead of students undertaking the other majors (Social studies or Language) as they learn more advanced mathematics. Given that the national average scores for mathematics were 42.87 (Year 9) and 37.23 (for Year 12 Science students), the performance of students from East Java was slightly above the national average score for mathematics. Thus, students from East Java were expected to represent average students in Indonesia.

Additionally, the selection process looked at students’ performance in Surabaya and Jombang compared to the other 36 cities in East Java (see Appendix H). Surabaya was in the third rank for Year 9 students (average score 56.30) and in the second rank for Year 12 students (average score 48.61). This indicated that Surabaya was among the top-performing cities in mathematics. In contrast, Jombang was in the 18th rank for Year 9 students (average score 47.96) and in the 19th rank for Year 12 students (average score 41.08). This indicated that students’ performance in Jombang was around the provincial average score. Therefore, participants from Surabaya and Jombang were expected to sufficiently represent the diversity of city performances in the UN 2019 test.

After justifying the city selection, the schools were identified. At the time of this study, there were 379 junior high schools and 135 senior high schools in Surabaya and 252 junior high schools and 68 senior high schools in Jombang, excluding vocational schools. Those schools were then classified based on the school type (under MoEC-RT or MoRA) and school status (public or private). This classification facilitated selection, which was based on schools’ average scores for mathematics in the UN. As a result, 16 schools were selected: eight junior high schools (for Year 9) and eight senior high schools (for Year 12). Of the eight schools in each grade level, four were in Jombang and four were in Surabaya. All the schools’ average mathematics scores were above or slightly above their city’s average. Schools in the lower rank were not selected; instead, students with low level of knowledge from the selected schools were involved.

Finally, six students from each of the 16 schools were selected, taking care to represent different genders and levels of knowledge. Particularly, two students were selected to represent each level of knowledge (low, medium and high) based on the judgement of their mathematics teachers. Additionally, the six students from each school represent the same number of boys and girls. Table 5.1 presents the distribution of the 96 participants, divided into Years 9 and 12. Further, a four-character code was created for each student rather than using a pseudonym to facilitate identification in the later stage (see Appendix I). The first character is a letter that refers to the grade level (‘A’ is for Year 9 and ‘B’ is for Year 12), the second and third characters uniquely identify the student, and the fourth character is a letter referring to the city (‘J’ is for Jombang and ‘S’ is for Surabaya). For example, the codes for the six Year 9 students from the public school under the auspices of MoRA in Jombang are A01J–A06J, whereas the codes for the six Year 12 students from the private school under the auspices of MoEC-RT in Surabaya are B19S–A24S. An ethical approval for this study was obtained from the University of Canberra Human Research Ethics Committee (see the approval from ethics committee in Appendix J).

Table 5.1

Distribution of Participants

	Jombang				Surabaya			
	MoRA		MoEC-RT		MoRA		MoEC-RT	
	public	private	public	private	public	private	public	private
Year 9	6	6	6	6	6	6	6	6
Year 12	6	6	6	6	6	6	6	6

5.4 Data Collection

5.4.1 SL Test

The SL test was administered in the first semester of Years 9 and 12, from November to December 2019 (see Table 5.2). All 96 participants from Surabaya and Jombang participated in the test, which was conducted in their own schools and overseen by the researcher. These consenting students voluntarily participated in the SL test and knew that they could withdraw at any time. During the test, the participants were under test conditions. They had 120 minutes to complete the test’s 10 items (see again Section 4.4 and Appendix F for the details of the 10-item instrument). However, many students finished within 90 minutes.

Table 5.2

Test Schedule

Test schedule	School	Year	City
November 2019	School I	9	Jombang
	School II	12	Jombang
	School III	12	Jombang
	School IV	12	Jombang
	School V	9	Jombang
	School VI	9	Jombang
	School VII	9	Jombang
	School VIII	9	Surabaya
	School IX	12	Jombang
	School X	12	Surabaya
	School XI	9	Surabaya
December 2019	School XII	9	Surabaya
	School XIII	12	Surabaya
	School XIV	12	Surabaya
	School XV	9	Surabaya
	School XVI	12	Surabaya

5.4.2 Interviews

Interviews were conducted with 25% of the tested students, that is, 24 students (see Table 5.3). The students were selected from the eight schools that allowed their students to be interviewed, and their names (pseudonyms) appear in Table 5.3. Pseudonyms were favoured over four-character codes to facilitate the use of their written responses in the study's findings. The students were selected immediately after each test and were chosen to represent a variety of written responses indicating different levels of performance. They were informed that they could withdraw at any time without getting penalised. The timing of the interviews was conducted at the students' schools depending on the students' and schools' availability.

Table 5.3

Interviewed Students

Year	Pseudonym	Gender	School type	School status	City of origin
Year 9	Ayu	Girl	MoRA	Public	Surabaya
	Budi	Boy	MoRA	Private	Jombang
	Cakra	Boy	MoRA	Public	Surabaya
	Dani	Boy	MoRA	Private	Jombang
	Ester	Girl	MoEC-RT	Public	Jombang
	Farah	Girl	MoRA	Private	Surabaya
	Galang	Boy	MoRA	Private	Surabaya
	Hannah	Girl	MoEC-RT	Public	Jombang
	Inggrid	Girl	MoRA	Private	Jombang
	Jessica	Girl	MoRA	Private	Surabaya
	Komar	Boy	MoRA	Public	Surabaya
	Luhut	Boy	MoEC-RT	Public	Jombang
Year 12	Maulana	Boy	MoRA	Private	Jombang
	Noval	Boy	MoRA	Private	Surabaya
	Oemi	Girl	MoRA	Private	Jombang
	Putra	Boy	MoRA	Private	Jombang
	Qiqi	Girl	MoEC-RT	Public	Surabaya
	Ridwan	Boy	MoEC-RT	Public	Surabaya
	Susi	Girl	MoRA	Private	Surabaya
	Thomas	Boy	MoRA	Private	Surabaya
	Ucok	Boy	MoEC-RT	Public	Jombang
	Vanes	Girl	MoEC-RT	Public	Jombang
	Wafiq	Boy	MoEC-RT	Public	Jombang
	Xavier	Boy	MoEC-RT	Public	Surabaya

Note. MoRA is The Ministry of Religious Affairs; MoEC-RT is The Ministry of Education, Culture, Research and Technology.

The 24 interviewed students were also selected to represent diverse backgrounds (grade levels, genders, school types, school statuses and cities). They were interviewed individually by the researcher for about an hour to verbally replicate and recall their thinking during the test. All interviews were video recorded. The lengths of time allocated to each item or each student were not necessarily the same. The way the students verbally communicated their thinking in response to the interviewer's follow-up questions determined the duration of the interview. The interview protocol was explained in Chapter 4.

5.5 Data Analysis

This section explains the overall analysis process used to address the four main research questions. Section 5.5.1 describes the analysis undertaken on the students' written responses, revealing the SL levels they achieved and their levels for the four skills (interpreting, communicating, evaluating and decision-making) and three components (text and context, representation and statistical-mathematical knowledge). The statistical analyses that were performed are also explained in this section.

Section 5.5.2 describes the analysis undertaken on the students' written responses and interviews to reveal the challenges they encountered and their understandings of three components when responding to the information containing statistics. The analysis employed the Constant Comparison Method (CCM). This analysis reveals students' challenges in and understandings of the three components when they interpret, communicate, evaluate and make decisions.

5.5.1 Quantitative Data Analysis: Students' SL Levels

The first research question addressed in this section is: 'What levels of SL do Indonesian high school students possess?' To answer this, a double coding procedure was employed. The main procedure in double coding is to have two or more coders code data independently then discuss it to obtain a consensus (Jones et al., 2000). Double coding was initially introduced by Miles and Huberman (1994) and has been employed in the field of statistics education by Jones et al. (2000) and Mooney (2002) to determine students' levels of statistical thinking. In the present study, three coders were involved to determine the SL levels possessed by students, by analysing their written responses. The three coders were two trained coders and the researcher. The two trained coders were mathematics lecturers; one holds a doctoral degree with a thesis on Indonesian high school students' SL, and another one holds a master's degree with a thesis on mathematics education using design research. This analysis resulted in the classification of students into each of the six hierarchical levels for the Year 9 and Year 12 students. The data covers the students' SL levels, skill levels, component levels and item component levels.

Following the double coding, a series of Mann-Whitney *U* tests were performed to answer the second research question: 'Are there any significant differences in Indonesian high school students' SL based on their demographic backgrounds (i.e., grade level, gender, school type, school status or city of origin)?' The Mann-Whitney *U* tests were used to investigate whether there were differences in students from different backgrounds, including grade level (Year 9 or Year 12), gender (boy or girl), school type (MoRA or MoEC-RT), school status (public or private) or city of origin (Jombang or Surabaya). This second research question suggested that all the statistical data analyses should be run within the SL assessment framework, in which SL is comprised of four response skills and three components.

In summary, the order of the various analyses performed on the students' written works was the double coding analysis then the Mann-Whitney U test. Each analysis is detailed below.

5.5.1.1 Double Coding

The double coding was conducted on the written responses of the 96 students by the three coders (two trained coders and the researcher). These students' personal information was non-identifiable and only the researcher who had access to it. Only eight of the 10 tested items were analysed. Two of the four items assessing the interpreting skill were excluded from the analysis based on the fact that the other skills (communicating, evaluating, and decision-making) were also represented by only two items involving one type of representation. With the two items removed, each skill then involved only one representation: a line graph for interpreting, a bar graph for communicating, a bar graph for evaluating and a table for decision-making. In this study, the double coding involved five stages: group coding, individual coding, inter-rater reliability check, consensus coding and deciding students' SL levels.

5.5.1.1.1 Group Coding

The group coding was a preliminary trial and involved three coders. The three coders encoded the written responses of 25% of the sample (equivalent to 12 students in Year 9 and 12 students in Year 12). These students were the interviewed students and were purposefully selected to represent students with diverse levels of knowledge and responses. The coders assigned the numerical code that best represented the students' understanding at different levels of knowledge and comprehension.

The group coding process began by assigning a code for each of the three components contributing to each student's SL: text and context, representation and statistical-mathematical knowledge. The numerical code represents the student's level and ranges from 1 (idiosyncratic), 2 (informal) and 3 (inconsistent) to 4 (consistent non-critical), 5 (critical) and 6 (critical mathematical). Prior to this coding process, component-based item descriptors were developed for each of the six levels (see Section 4.3.3). Eventually, the group coding resulted in a code for each of the three components tested by a particular item, to represent each student's level in each component.

The three coders interpreted the 24 students' written responses through a WhatsApp phone conference. Each of the three coders had a printed copy of the students' written answers with them. During the interpretation process, one coder presented a reasonable interpretation of a student's response regarding 1) what the student's written responses represented, 2) how it could be interpreted and 3) what numerical code should be assigned for each component of SL. The other two coders then gave their comments on the first coder's interpretation and the three numerical codes suggested. In the event of a disagreement, the three coders discussed their reasoning until either consensus was reached (Mooney, 2002) or new descriptors were generated if they were not present in the component-based item descriptors. This process was repeated for all 24 students, and each coder took turns at being the first to provide a reasonable interpretation and suggest the codes that should be assigned. The items were coded one at a time. Once coding for 24 students was completed on one item, the coding moved to the individual coding (see Section 5.5.1.1.2) to code the remaining 72 students for that item. Once coding on the one item for all students was completed, group coding recommenced for the second item. This cycle was repeated until the eighth item had been coded.

5.5.1.1.2 Individual Coding

Having practised by coding the written responses of 25% of the students, the three coders each coded the responses of the remaining 75% of students (72 students, in Years 9 and 12) independently. Each of the three coders applied the same coding techniques as employed during the group coding stage. If component-based item descriptors still could not be applied to particular students' responses, those responses were re-examined thoroughly until the closest corresponding descriptor could be identified (Mooney, 2002), by individual coder. Otherwise, a record was made by the individual coder, to be discussed during the consensus coding (see Section 5.5.1.1.4).

The individual coding was conducted in two rounds for each item. In the first round, the codes from the three coders were recorded in a table. From this table, the students with adjacent agreement (one-level difference), non-adjacent agreement (two-level difference or more) and complete disagreement (no identical codes from three coders) were identified and marked for further recoding. The second round of individual coding then focused only on those students. During this individual recoding, however, each coder knew only their own previous codes and not those from the other two coders. Finally, the results from the individual recoding were recorded and the inter-rater reliability check was conducted.

5.5.1.1.3 Inter-Rater Reliability Check

Following the individual coding and recoding, statistical data analyses were conducted to check the strength of agreement between the three coders. Kendall's *W* (Laerd Statistics, 2016) determined if there was agreement between the three coders on the codes, they assigned to each of the three components. Kendall's *W* was chosen because there were three coders, 72 students and three variables (the three components contributing to students' SL) and the students' level in each component was ordinal data. The results showed that the three coders significantly agreed in the levels they provided ($p < .0005$), with $W \geq .814$ and the majority of values above .900, which is considered very strong agreement. Table 5.4 summarises the Kendall's *W* results.

Table 5.4

Kendall's W Results for All Three Components of the Eight Assessment Items

Item	Skill	Kendall's *W* value		
		Text and context	Represen tation	Statistical- mathematical
The Production Mean	Interpreting	.903	.922	.923
The Most Production	Interpreting	.956	.939	.939
YouTube Viewers	Communicating	.814	.901	.913
Domestic Waste	Communicating	.911	.941	.820
Math Scores	Evaluating	.926	.881	.917
The Employees	Evaluating	.922	.912	.936
The 100-Metre Race	Decision-making	.948	.932	.932
Which Motorcycle?	Decision-making	.872	.898	.886

5.5.1.1.4 Consensus Coding

In cases where the three coders disagreed at the individual coding stage, the disputed codes were discussed during consensus coding. The three coders conducted consensus coding through a WhatsApp phone conference. Each consensus coding session always started by listening to the reasoning of the person coding differently. After hearing that coder's reasoning, the other two coders gave responses and discussion followed to obtain consensus. In the case of complete disagreement (no identical codes from three coders), any coder could voluntarily present their reasoning; this would be followed by a consensus discussion.

5.5.1.1.5 Deciding Students' SL Levels

The numerical codes that were resulted from the previous stages of coding indicated each participant's apparent level of knowledge of each of the three SL components (text and context, representation and statistical-mathematical knowledge). The median of the SL components' codes further characterised the code (i.e., level) for each component, for each item, for each skill and for overall SL. The median was chosen instead of the mean as it is the recommended measure of central tendency for ordinal data (Boone & Boone, 2012; Harpe, 2015; Joshi et al., 2015; Stevens, 1946). In cases where the median was halfway between two levels, it was rounded down to ensure that participants' responses were coded to the nearest corresponding descriptors (see Table 5.5). This rounding followed Mooney (2002), who rounded down the mean when it was halfway between two levels when determining students' statistical thinking levels using various constructs. For example, if a participant's knowledge level for text and context based on two interpreting skill items was coded as 4 and 4, for representation as 4 and 5, and for statistical-mathematical knowledge as 5 and 5, the median would be 4.5 = [(4 + 5) ÷2]. The median would then be rounded down to the lower level, resulting in the participant receiving a Level 4 for the interpreting skill (i.e., consistent non-critical). This process was applied for each item, each skill, each component and the overall SL. The participant's overall SL level was derived from the median of the components' codes they obtained in all items.

Table 5.5

Rounding of Median

Median	Code
1.0 ≤ median ≤ 1.5	1
1.5 < median ≤ 2.5	2
2.5 < median ≤ 3.5	3
3.5 < median ≤ 4.5	4
4.5 < median ≤ 5.5	5
5.5 < median ≤ 6.0	6

5.5.1.2 Descriptive Statistics

As this study was intended to cross-sectionally compare the SL levels of Year 9 and Year 12 students, the levels derived from the above double coding process were presented accordingly. The SL level of students in both grade levels were presented to allow for the initial identification of differences. Additionally, it allowed for the explanation and comparison of students' levels in four skills (interpreting, communicating, evaluating and decision-making) and three components (textual and contextual understanding, graphical competence and statistical-mathematical knowledge), either within the same grade level or across both grade levels.

In reporting the students' SL levels, the distribution of levels achieved by the students was presented using tables. Tables were used rather than graphs to easily compare multiple variables (such as grade level, four skills, three components and six hierarchical levels). Moreover, tables were easy to use for comparison when the six levels were classified into two groups: the lower group, consisting of Levels 1 (idiosyncratic) to 3 (inconsistent), and the upper group, consisting of Levels 4 (consistent non-critical) to 6 (critical mathematical). This classification considers the lower group to be those students encountering challenges in comprehending the three components, and the upper group to be those demonstrating appropriate and critical understanding of the three components when responding to data-based information. A number of comparisons were then generated from this distribution, including the percentage of students in the lower and upper group within the same grade level, the percentage of students in the upper group from both grade levels and the level with the highest percentage of the six hierarchical levels between two grade levels.

Furthermore, the proportions of students in the upper group were examined to assess both component difficulty and item component difficulty in relation to item difficulty. The difficulty of a component for students increases as the proportion of students in the upper group decreases. Table 5.6 presents the categories used in this study for interpreting levels of difficulty. Based on this table, the difficulty of each component across different grade levels could be compared. This analysis was intended to identify which components that presented the most challenges to students when responding to data-based information. Similarly, understanding the difficulty of an item's components helped to determine which specific components posed challenges for students in particular item and whether that item was more or less challenging compared to others.

Table 5.6

Level of Difficulty Interpretation

Percentage of students in the upper group	Level of Difficulty
0% – 20%	Very difficult
21% – 40%	Difficult
41% – 60%	Moderate
61% – 80%	Easy
81% – 100%	Very easy

5.5.1.3 Statistical Data Analysis: Identifying Differences Between Grade Levels

The next step was to determine whether there were differences in students' performance dependent on their demographic backgrounds. Mann-Whitney *U* tests were then performed, as the students' level is ordinal data (Laerd Statistics, 2015). Mann-Whitney *U* tests were run using SPSS (Version 27 for Windows) and double-checked with Jamovi (Version 2.3.16 for Windows).

In the first Mann-Whitney *U* test, the dependent variables were the SL levels and those of the four response skills (interpreting, communicating, evaluating and decision-making), while the independent variable was the grade level (Year 9 or Year 12). Before presenting the results in Chapter 6, it is important to check the four preliminary assumptions for Mann-Whitney *U* test in this chapter. The preliminary assumption checking for this first Mann-Whitney *U* test (see Table 5.7) revealed that the first three assumptions about the study design had been met—guided by Laerd Statistics (2015). Following on, the fourth assumption, about the differences in the distributions of SL and skill level for both grade levels, had been violated, as shown in the histogram produced by SPSS (see Figure 5.1). These distributions of SL and skill levels for both grade levels have different shapes. As a result, a Mann-Whitney *U* test was used to determine whether there were differences in the mean ranks of the grade levels using p < .05 as the level of significance; in other words, mean rank was used instead of median in the reporting.

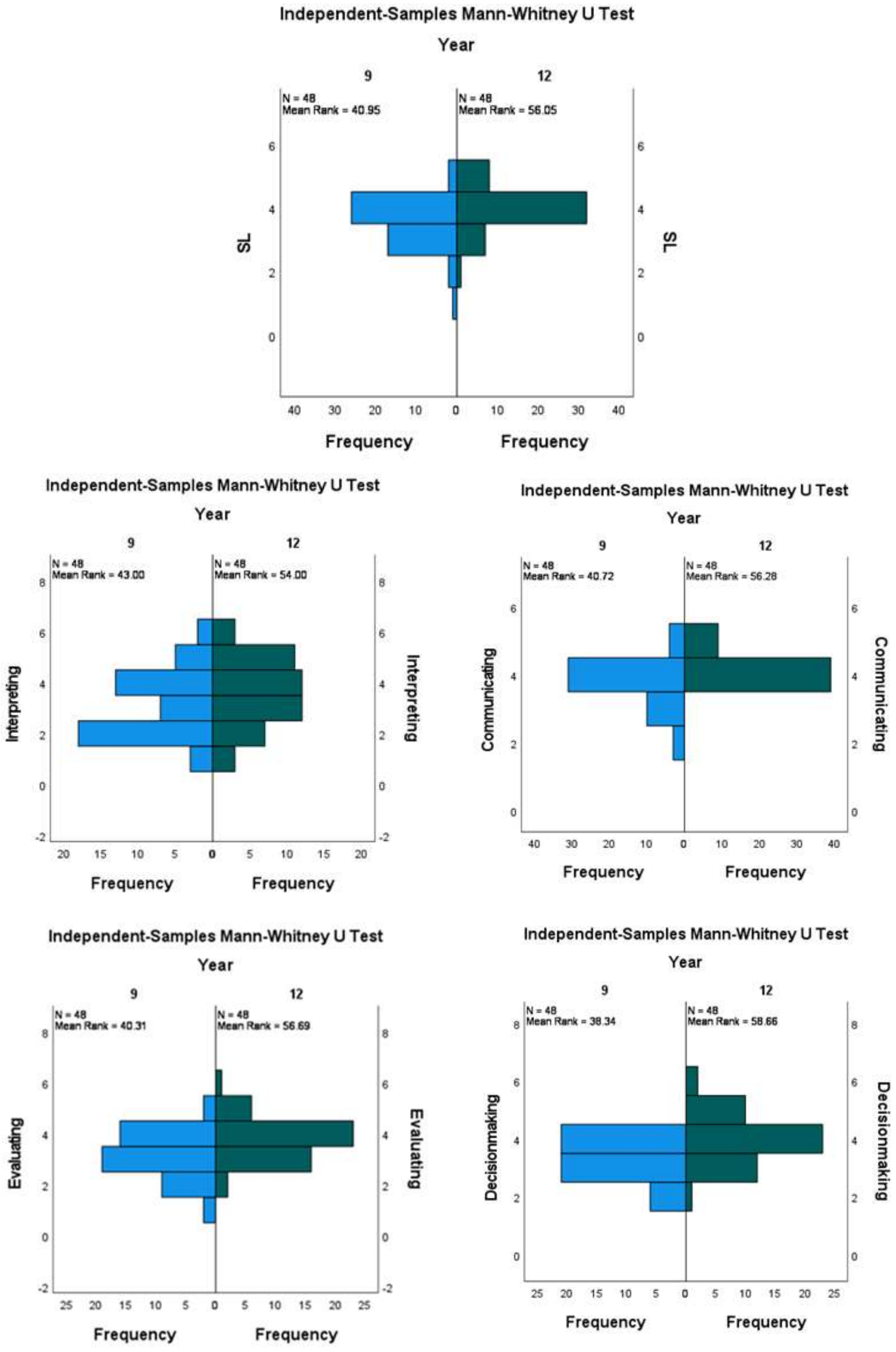
Table 5.7

Results of the First Three Assumption Checks for the Mann-Whitney U Test

Assumption	Checking	Result
Each dependent variable is continuous or ordinal.	Each of the dependent variables (SL level, skill level and item level) are ordinal data.	Met
There is one independent variable that categorises two independent groups.	The independent variable is grade level, which categorises two independent groups (Year 9 and Year 12).	Met
Independence of observations: this means there is no relationship between the observations made for each independent variable group or between the groups.	None of the participants in Year 9 were also in Year 12.	Met

Figure 5.1

Differences in the Distribution of SL Levels and Skill Levels Between Years 9 and 12



5.5.1.4 Statistical Data Analysis: Identifying Differences Based on Participants' Demographic Backgrounds

The researcher was also interested in determining whether there were differences in the students' SL levels based on their demographic. As described in Section 5.3, participants were selected to represent various demographics, namely gender, school type, school status and city of origin. As with grade level, Mann-Whitney *U* tests were used to investigate the differences between participants from different backgrounds. The four assumptions for all Mann-Whitney *U* tests were checked and similar results were found to those for the Mann-Whitney *U* tests for grade level. Consequently, the mean rank was used instead of the median in determining whether there was a statistically significant difference between the two groups being compared, using p < .05 as the level of significance.

5.5.2 Qualitative Data Analysis: Students' SL-Related Challenges and Understandings

This section presents the qualitative data analysis to address the third and fourth research questions of this study: 'How do the challenges students encounter in comprehending the three components of SL affect their abilities to respond to statistical information?' and 'How do students' understandings of the three components of SL influence their abilities to respond to statistical information?' The analysis was conducted on the written responses of 24 students, supported by their interviews.

To reveal the students' challenges and understandings when responding to data-based information, a CCM was grounded (Glaser & Strauss, 2017) based on students' written response and their interview. The written and oral responses of 24 students were used to shed light on the students' challenges and understandings. A systematic, yet flexible, method for studying the emerging data through constant comparisons of data and categories rather than a priori theory was conducted in accordance with the grounded theory guidelines (Case & Jacobbe, 2018). This study modified the purposeful approach of the CCM offered by Boeije

(2002). Boeije’s approach consisted of five steps and aimed to systematise the process of analysis and to increase the traceability. When describing how the approach was implemented, he clarified some issues related to the subject of the comparison, phases of the comparison, reason for the comparison and results of the comparison. In the present study, four stages of CCM were applied to address the same issues and ensure that all data in the data set were compared. The four stages are explained below using examples from the challenges students faced in and their understandings of the interpreting skill. Analysis of the students’ challenges and understandings for the other three skills (communicating, evaluating and decision-making) followed these four stages.

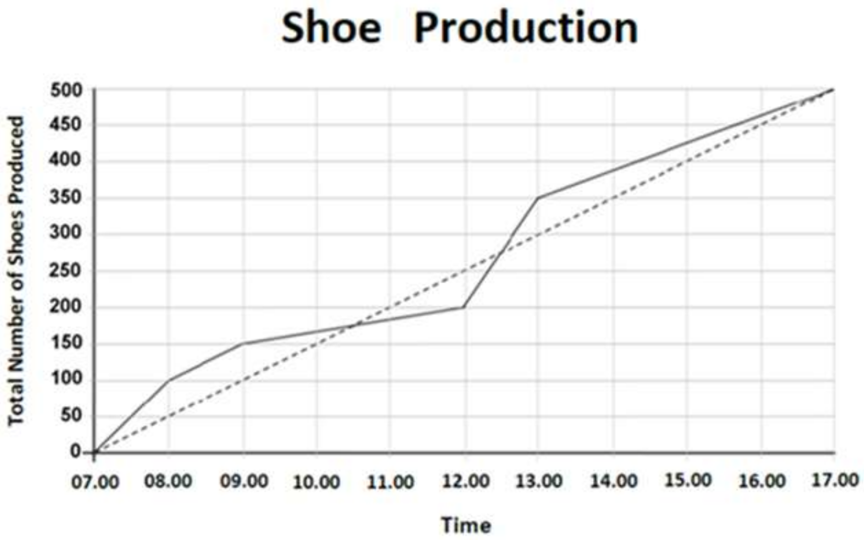
5.5.2.1 Comparison Within a Student’s Written Response for a Single Item

In this first comparison, every student’s response for a single item was analysed to determine what the student had written. This internal comparison was grounded based on the student’s written responses and aimed to identify what challenges or understandings a student experienced when solving one particular item. Three coders were involved, and this first comparison was conducted concurrently with the group coding, as explained in Section 5.5.1.1.1. The three coders first examined the student’s challenges and understandings by comparing different parts of their written response (writing, drawing, calculation, sketch or any other signs). The student’s challenges or understandings were examined relative to the three SL components. For this purpose, some important questions guided the three coders performing comparison analysis in this first phase, such as: ‘What did this student do to solve the problem?’ and ‘Were the student’s challenges or understandings reflected in their written responses?’ This first comparison generated a consensus on each student’s challenges and/or understandings for a single item relative to the three SL components.

To illustrate this process, Figure 5.2 shows The Production Mean item and one written response from Dani. The Production Mean item was designed to assess the students’ interpreting skills. This item was set under the context of shoe production. This item asked students: ‘What was the mean number of shoes produced per hour? Explain how you got it!’ This was intended to assess students’ comprehension of the production mean as a measure of central tendency from a line graph. For this item, students were required to comprehend that the solid line displays the raw data gathered at certain times during a particular day, while the dotted line exhibits the processed data as if in a constant increase. Students were expected to critically justify that the production mean is shown by the hourly constant increase of the dotted line.

Figure 5.2

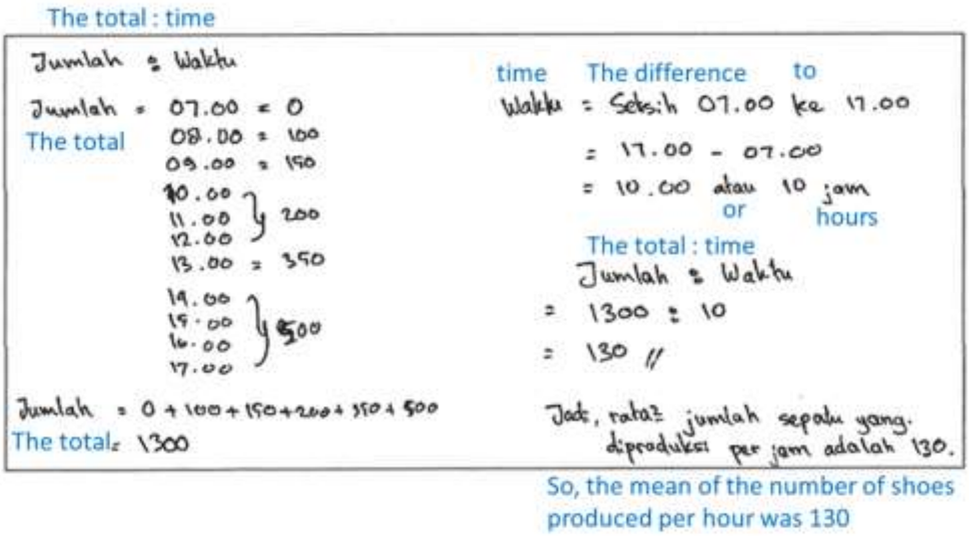
The Production Mean Item and Dani’s Written Response for Comparison within a Student’s Written Response



The solid line (——) on graph shows the number of shoes produced by a home industry during a particular day.

The dotted line (- - - -) shows what the total number of shoes produced would be if the rate of production were constant.

Figure 5.2 (continued)



The total : time

Jumlah : Waktu

Jumlah = 07.00 = 0

The total

08.00 = 100

09.00 = 150

10.00
11.00
12.00 } 200

13.00 = 350

14.00
15.00
16.00
17.00 } 500

Jumlah = 0 + 100 + 150 + 200 + 350 + 500

The total = 1300

time The difference to

Waktu = Selisih 07.00 ke 17.00

= 17.00 - 07.00

= 10.00 atau 10 jam

or hours

The total : time

Jumlah : Waktu

= 1300 : 10

= 130 //

Jadi, rata2 jumlah sepatu yang diproduksi per jam adalah 130.

So, the mean of the number of shoes produced per hour was 130

In discussing Dani’s response in Figure 5.2, one of the three coders began by interpreting what challenges Dani had in solving The Production Mean item. From Dani’s response, it was interpreted that he attempted to find the total number of shoes produced (1,300) and the total time of production (10 hours). After finding both, Dani divided 1,300 by 10, resulting in 130. At the end of his writing, he concluded, ‘So, the mean number of shoes produced per hour is 130’. The three coders agreed that Dani encountered challenges in comprehending the three components of SL, as presented in Table 5.8.

Table 5.8

Challenges Dani Encountered in The Production Mean Item

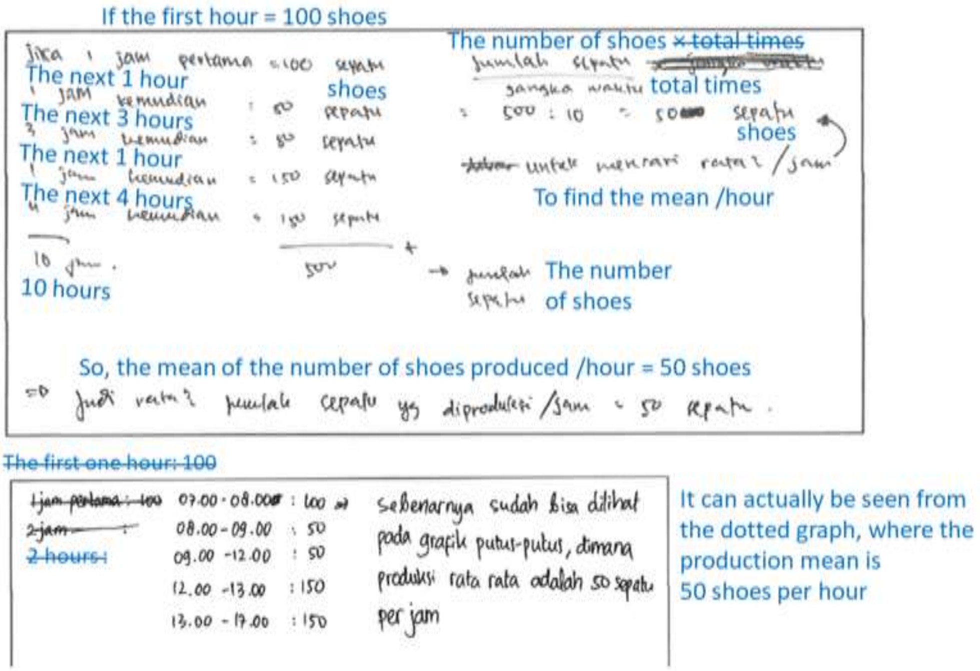
Item	Text and context	Representation	Statistical-mathematical knowledge
The Production Mean	Dani focused merely on the solid line and ignored the dotted line. Dani did not understand the idea of ‘increase’.	Dani failed to notice that the data was discrete and cumulative. Dani failed to notice data points, data increments and constant increase.	Dani understood the procedure to find a mean from formula. Dani performed a correct number operation, especially addition and division. Despite Dani’s number operation being correct, the mean was incorrect.

5.5.2.2 Comparison Between Students’ Written Responses for a Single Item

The second comparison was between the students’ written responses for the same item. This comparison aimed to find commonalities or differences between students’ challenges or understandings in regard to the SL component. To achieve this aim, some important questions guided the three coders performing this comparison analysis. Some of those questions were: ‘What are the commonalities and differences between students’ responses?’ and ‘If the challenges or understandings differ, what makes it different?’ This comparison generated various challenges and/or understandings for a single item. To illustrate this process, Figure 5.3 shows the written responses from two students on The Production Mean item.

Figure 5.3

Comparing Inggrid's and Xavier's Written Responses for The Production Mean Item



If the first hour = 100 shoes

Jika 1 jam pertama = 100 sepatu

The next 1 hour shoes

1 JAM kemudian : 50 sepatu

The next 3 hours

3 jam kemudian : 50 sepatu

The next 1 hour

1 jam kemudian : 150 sepatu

The next 4 hours

4 jam kemudian : 150 sepatu

10 jam. 500 +

10 hours

→ jumlah sepatu The number of shoes

The number of shoes ~~× total times~~

jumlah sepatu ~~jangka waktu~~

jangka waktu total times

= 500 : 10 = 50 sepatu shoes

~~dalam~~ untuk mencari rata2 /jam

To find the mean /hour

So, the mean of the number of shoes produced /hour = 50 shoes

=> Jadi rata2 jumlah sepatu yg diproduksi /jam = 50 sepatu.

~~The first one hour: 100~~

~~1 jam pertama : 100~~ 07.00 - 08.00 : 100 →

~~2 jam~~ 08.00 - 09.00 : 50

~~2 hours :~~ 09.00 - 12.00 : 50

12.00 - 13.00 : 150

13.00 - 17.00 : 150

Sebenarnya sudah bisa dilihat pada grafik putus-putus, dimana produksi rata rata adalah 50 sepatu per jam

It can actually be seen from the dotted graph, where the production mean is 50 shoes per hour

Note. The response at the top is from Inggrid, while the response at the bottom is from Xavier.

Both Inggrid and Xavier did not find any challenges in interpreting the information. Instead, they both showed understandings of the three components. It was interpreted that both students attempted to find the total number of shoes produced based on the data points. However, Inggrid then continued with the add-divide formula, while Xavier did not perform any calculation; instead, Xavier wrote a summary statement that justified the mean is 50 per hour as can be seen from the dotted line. As a result, Ingrid's understandings slightly differ to Xavier's understandings. Table 5.9 presents the understandings of the two students.

Table 5.9

Inggrid's and Xavier's Understandings in The Production Mean Item

Item	Text and context	Representation	Statistical-mathematical knowledge
The Production Mean	Both students understood that the two lines represented the same data: the solid line represented the raw data, and the dotted line represented the processed data. Both students understood the word 'constant' and the idea of 'increase'.	Both students understood the line graph convention (such as the label for x and y-axis). Both students knew that the data was discrete and cumulative. Both students knew data points, data increments and constant increase.	Inggrid knew the concept of mean and used the mean formula, while Xavier knew the concept of mean from both mean formula and graph. Inggrid correctly performed number operation, especially addition and division.

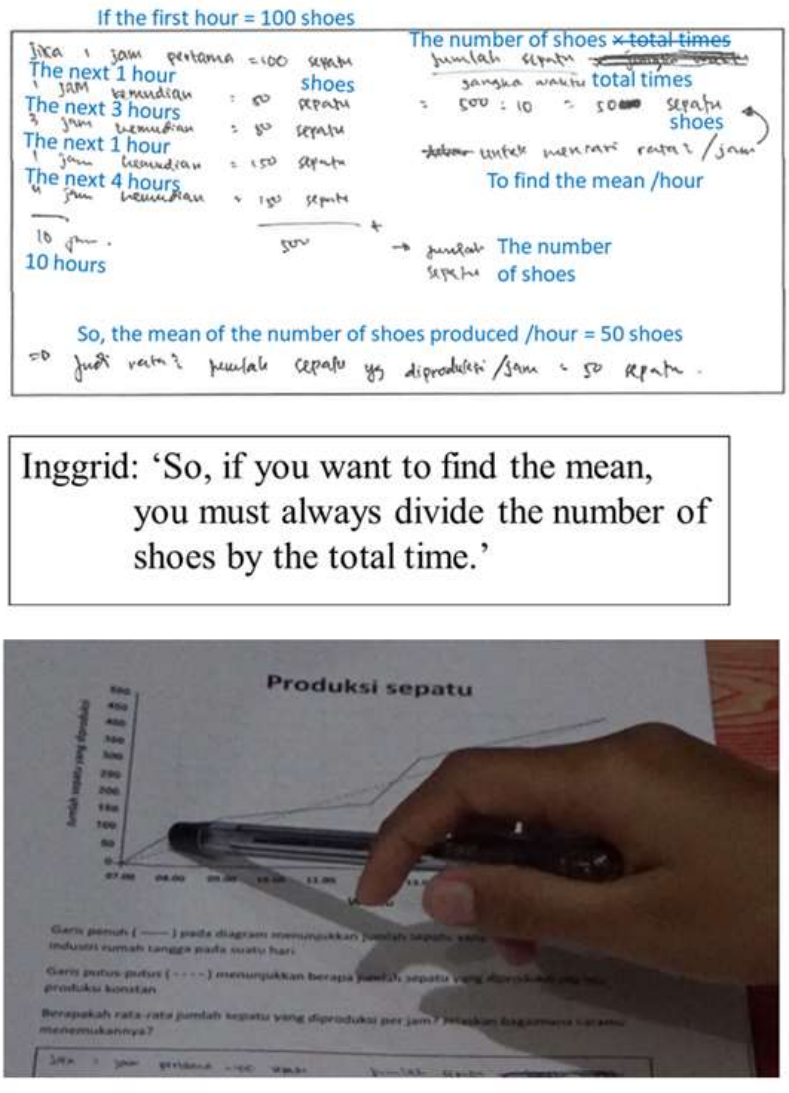
5.5.2.3 Comparison Between Students' Response and Interview

The third comparison was between a student's written response and interview for each item. This comparison was conducted by one coder (the researcher) and intended to double-check whether students' challenges and understandings identified in the previous two stages were adequate and appropriate. This comparison sought further evidence on students' challenges and/or understandings. Some important questions guided the researcher, including 'Were they the actual challenges or understandings that students encountered?', 'What was their thinking?', 'Does the interview support or contradict the written response?' and 'Is there new evidence from the interview that does not exist in the written response?' This comparison confirmed the various challenges and/or understandings for a single item.

The written responses from Inggrid and Xavier on The Production Mean item are used to illustrate this comparison process. Their gesture and oral response during interview were used to confirm their understandings. Figure 5.4 illustrates the confirmation of Inggrid's understandings, while Figure 5.5 illustrates the confirmation of Xavier's understandings. No new information appeared from Inggrid’s interview. Comparatively, it was found that Xavier also mentally calculated before concluding that the mean was the hourly increase in the dotted line.

Figure 5.4

Comparing Inggrid’s Written Response, Oral Explanation and Gesture During the Interview



If the first hour = 100 shoes

Jika 1 jam pertama = 100 sepatu

The next 1 hour

1 jam kemudian = 50 sepatu

The next 3 hours

3 jam kemudian = 80 sepatu

The next 1 hour

1 jam kemudian = 150 sepatu

The next 4 hours

4 jam kemudian = 120 sepatu

10 jam.

10 hours

500

→ jumlah sepatu The number of shoes

The number of shoes ~~× total times~~

jumlah sepatu ~~[illegible]~~ / jangka waktu total times

= 500 : 10 = 50 sepatu shoes

~~[illegible]~~ untuk mencari rata² /jam

To find the mean /hour

So, the mean of the number of shoes produced /hour = 50 shoes

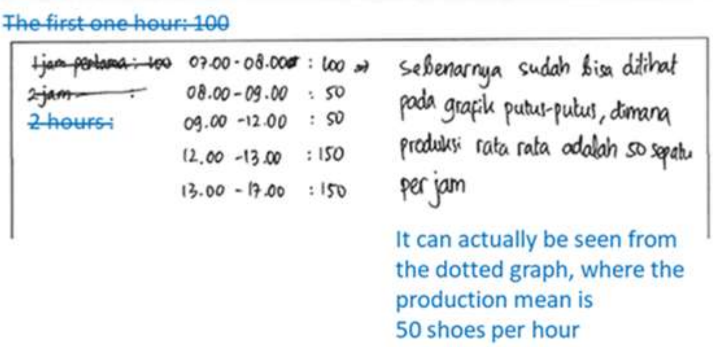
=> Jadi rata² jumlah sepatu yg diproduksi /jam = 50 sepatu.

Inggrid: ‘So, if you want to find the mean, you must always divide the number of shoes by the total time.’

Note. The figure at the top depicts Inggrid’s written response; the figure in the middle depicts Inggrid’s oral explanation; the figure at the bottom depicts Inggrid’s gesture during the interview.

Figure 5.5

Comparing Xavier’s Written Response, Oral Explanation and Gesture During the Interview



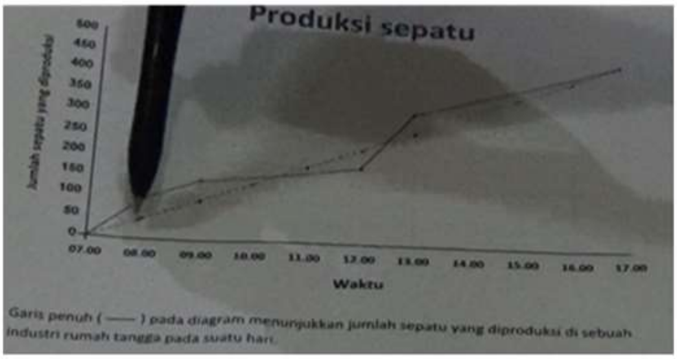
~~The first one hour: 100~~

~~1 jam pertama : 100~~ 07.00 - 08.00 : 100 →
~~2 jam~~ 08.00 - 09.00 : 50
~~2 hours :~~ 09.00 - 12.00 : 50
12.00 - 13.00 : 150
13.00 - 17.00 : 150

Sebenarnya sudah bisa dilihat pada grafik putus-putus, dimana produksi rata rata adalah 50 sepatu per jam

It can actually be seen from the dotted graph, where the production mean is 50 shoes per hour

> Xavier: ‘After reading the question, I tried to find the mean, and after that I concluded that the mean was there in the dotted line [pointing to the increment in the dotted line].’



Note. The figure at the top depicts Xavier’s written response; the figure in the middle depicts Xavier’s oral explanation; the figure at the bottom depicts Xavier’s gesture during the interview.

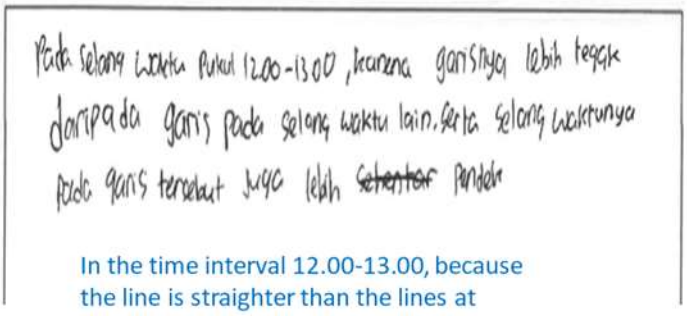
5.5.2.4 Comparison of Students’ Responses from Different Items in the Same Skill

The final comparison was between a student’s challenges or understandings from two items assessing the same skill. This comparison was also conducted by one coder (the researcher) and intended to identify the students’ challenges and/or understandings for the four skills (interpreting, communicating, evaluating and decision-making) in regard to the three components (text and context, representation and statistical-mathematical knowledge). Some important questions guided the researcher, such as: ‘What challenges or understandings from the two items can be generalised?' This comparison revealed the students' challenges and/or understandings for a certain skill.

The written responses from Ucok on the second interpreting skill item, The Most Production item, is used to illustrate this comparison. The Most Production item was set under the same context as the Production Mean item. This item asked students, 'During which time interval were the most shoes produced? Explain and relate your answer with the shoe production graph.' It aimed to assess students' ability to locate the time interval in which the most shoes were produced within the line graph. Students were expected to reason that the most production occurred between the two data points connected by a line with the sharpest slope. Figure 5.6 illustrates Ucok's written response, gesture and interview showing his understandings. Ucok's understandings of the three components of The Most Production item were then compared to a collection of students' understandings for The Production Mean item. This comparison showed there were similarities between the sharpest slope of Ucok and the constant increase of Xavier. Therefore, both were classified into understandings of graph convention that refer to the component of representation.

Figure 5.6

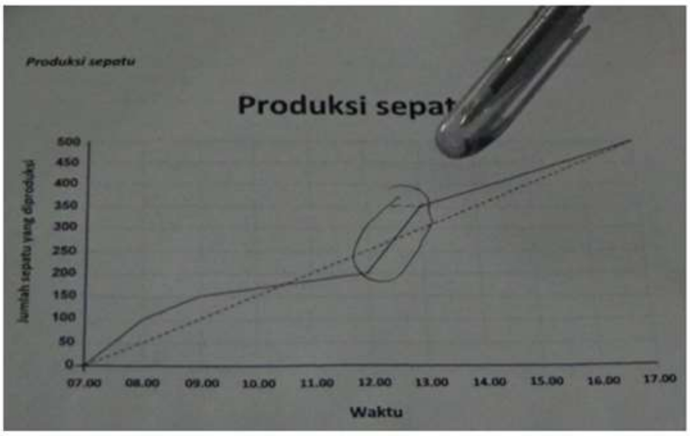
Comparing Ucok’s Written Response, Oral Explanation and Gesture During the Interview



Pada selang waktu Pukul 12.00-13.00, karena garisnya lebih tegak daripada garis pada selang waktu lain. serta selang waktunya pada garis tersebut juga lebih ~~Sebentar~~ Pendek

In the time interval 12.00-13.00, because the line is straighter than the lines at other time intervals. In addition, the time interval on this line is shorter

Ucok: ‘The line with the sharpest slope almost upright … the sharper the line the lesser time …. the more shoes produced.’



Note. The figure at the top depicts Ucok’s written response; the figure in the middle depicts Ucok’s oral explanation; the figure at the bottom depicts Ucok’s gesture during the interview.

This comparison concluded the four stages of CCM. In reporting the results, the students’ challenges on the three SL components were presented in relation to the four skills assessed. Similarly, the students’ understandings of the three SL components related to the four skills assessed were revealed.

5.6 Chapter Summary

This chapter has described this study’s cross-sectional design and the methods used to collect and analyse the cross-sectional data. This study implemented an explanatory sequential design that began with a quantitative component followed by a qualitative component. A test was administered with a 10-item instrument, and a follow-up interview was conducted to clarify students’ thoughts during the test. For the quantitative component, various analyses were performed on students’ written works (double coding, descriptive statistics and Mann-Whitney U test). The four stages of CCM were employed for the qualitative component.

The results of the data analysis are presented in Chapters 6 and 7. Chapter 6 presents the results of the quantitative analyses pertaining to the first two research questions: students’ SL levels and whether students’ SL levels develop with age and differ by other demographic backgrounds (student gender, school type, school status and city of origin). Chapter 7 provides the results of the qualitative data analysis pertaining to the third and fourth research questions, demonstrating students’ challenges in and understandings of the three components of SL.

Chapter 6: Students' SL

This chapter presents the findings from the analyses of students' written responses to data-based items and the discussion of some key findings. Section 6.1 presents the findings to this study's first research question: 'What levels of SL do Indonesian high school students possess?' Section 6.2 presents the findings of statistical analyses to answer the second research question: 'Are there any significant differences in Indonesian high school students' SL based on their demographic backgrounds (i.e., grade level, gender, school type, school status or city of origin)?' Some key findings are discussed in Section 6.3. Section 6.4 summarises this chapter, serving as the basis for Chapter 7, which presents the qualitative findings.

6.1 Students' SL Levels

This section presents the descriptive statistics of students' SL levels based on the assessment framework. Section 6.1.1 presents the frequency distribution of the Year 9 and Year 12 students' overall SL levels across the six hierarchical levels. This distribution is then used to initially identify Year 12 students' development regarding the SL. The six levels were classified into two groups: the lower (Levels 1–3) and upper (Levels 4–6) groups, representing students with limited and advanced statistical thinking, respectively (see the example of students' responses in Appendix K). Section 6.1.2 presents the distributions of the levels achieved by students in both grade levels on each of the four response skills (interpreting, communicating, evaluating and decision-making). This distribution is used to initially identify if any progress has been made by Year 12 students in relation to the four SL skills. Using the same pattern of analysis, Section 6.1.3 presents the distributions of the levels achieved by students for the three components (text and context, representation and statistical-mathematical knowledge). This distribution is also used to determine the difficulty order (very easy, easy, moderate, difficult and very difficult [see Section 5.5.1.2]) of the item and three components for each item. The summary finding is presented in Section 6.1.4 to answer the first research question.

6.1.1 Distribution of Students' SL Levels

Table 6.1 highlights the different proportions of Year 9 and Year 12 students' performances in the lower and upper groups, facilitating within grade-level comparisons. While the number of Year 9 students in the lower and upper groups was fairly balanced, the number of Year 12 students in the upper group outnumbered their lower group counterparts by about five times. While 42% of Year 9 students demonstrated limited statistical thinking when responding to data-based information, a significant majority of Year 12 students (83%) exhibited appropriate or critical levels of statistical thinking.

Table 6.1

Distribution of Students' SL Levels Across the Hierarchy by Percentage

Year	Lower group			Upper group		
	L1	*L2*	*L3*	*L4*	*L5*	*L6*
9	2%	4%	36%	54%	4%	0%
12	0%	2%	15%	67%	16%	0%

Note. L1 signifies Level 1 etc.

The proportion also facilitates between grade-level comparisons, potentially indicating SL progression in high school. Notably, Level 3 (inconsistent) Year 9 students were more than double that of Year 12 students, and some Year 9 students only reached Level 2 (informal) and Level 1 (idiosyncratic). In the upper group, both Year 9 and Year 12 students achieved Level 4 (consistent non-critical) and 5 (critical), but Year 12 students showed a higher proportion in both levels. This suggests that high school students' SL may improve during their time in high school.

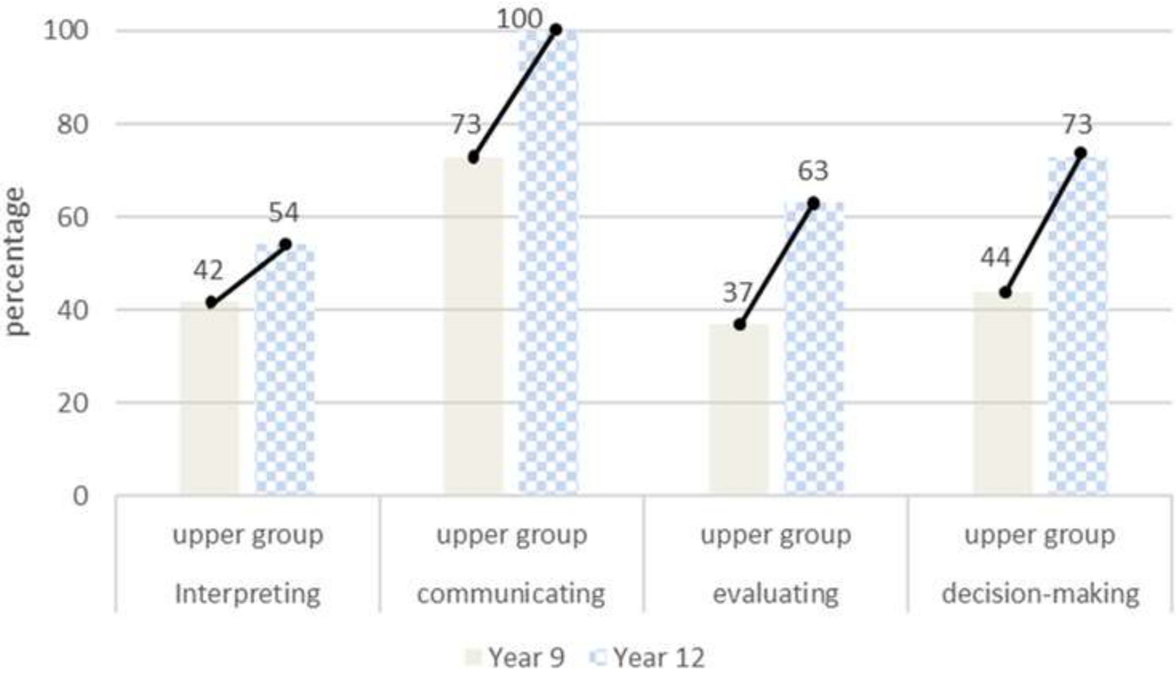
Nevertheless, both grade levels showed a similar pattern in SL levels, with Year 12 outperforming Year 9. The highest percentage of students in both grade levels achieved Level 4 (consistent non-critical). Notably, no student in either grade achieved the highest level (Level 6: critical mathematical). The inability of these Year 9 students to achieve Level 6 was initially predicted due to previous low performances of Indonesian 15-year-olds in PISA statistical problems. However, it is alarming that no Year 12 students achieved Level 6, as they are reaching the end of their secondary schooling. This prompted a deeper investigation into the levels of SL skills attained by these Year 12 students, in comparison to Year 9 students.

6.1.2 Distribution of Students' Levels in the Four Skills

As the SL comprises four skills (interpreting, communicating, evaluating and decision-making), the frequency distribution of students' levels in the four response skills is presented. Figure 6.1 presents the percentage of students in the upper group for both grade levels in four skills. Figure 6.1 also generates two main findings regarding the comparison of the performances of students across both grade levels. First, in the upper group, there were more Year 12 students than Year 9 students in all skills. This result indicates that SL skills might develop across high school, complementing the results of students' SL. Second, students in both grade levels performed better in the communicating skill than the other three skills. In this skill, all Year 12 students were able to demonstrate statistical thinking in their written responses, whereas around 73% of Year 9 students were able to do so.

Figure 6.1

Percentage of Upper Group Students in All Skills



Beyond the grade level comparison, the distribution of students across the six hierarchical levels offers further insights. Table 6.2 indicates that the highest and second-highest number of students performed either in Level 4 (consistent non-critical) or Level 3 (inconsistent) for nearly all the response skills, including in communicating. This suggests that most students were in between inconsistent and appropriate, albeit non-critical, data-based thinking. However, the Year 12 and Year 9 trends show differences. For Year 12, there was a marked achievement in Level 4 across all skills except interpreting, where students were evenly distributed between Levels 3 and 4. Unique to interpreting, Year 12 students spread across all six levels, contrasting the communicating skill which was limited to Levels 4 and 5. In evaluating and decision-making, Year 12 students displayed a similar pattern across five levels (2–6), with the highest percentage of students in Level 4 followed by Level 3.

Table 6.2

Distribution of Students' Skill Levels Across the Hierarchy by Percentage

	Year	Lower group			Upper group		
Skill		*L1*	*L2*	*L3*	*L4*	*L5*	*L6*
Interpreting	9	6%	37%	15%	27%	11%	4%
	12	6%	15%	25%	25%	23%	6%
Communicating	9	0%	6%	21%	65%	8%	0%
	12	0%	0%	0%	81%	19%	0%
Evaluating	9	4%	19%	40%	33%	4%	0%
	12	0%	4%	33%	48%	13%	2%
Decision-making	9	0%	12%	44%	44%	0%	0%
	12	0%	2%	25%	48%	21%	4%

Note. L1 signifies Level 1 etc.

Different to Year 12, the distribution for Year 9 students varies distinctly in each skill. In communicating, the highest number of students was in Level 4, which stands as the largest proportion across the four skills. While Levels 3 and 4 saw a similar count of students in decision-making, the highest and second-highest number of students were in Level 3 and Level 4 in evaluating. Interestingly, in interpreting, the highest number of Year 9 students was at Level 2 (informal), followed by Level 4 (consistent non-critical). This distinct distribution of Year 9 students in the interpreting skill warrants deeper analysis. A potential method could involve examining their achievement across components of interpreting items—compared to the other items—as detailed in the subsequent section.

6.1.3 Distribution of Students' Levels in the Three Components of SL

This section examines component difficulty across grade levels in addition to the distribution of student achievement in three components—text and context, representation and statistical-mathematical knowledge. The criteria of the difficulty level were previously presented in Chapter 5 (see again Table 5.6). By presenting the frequency distribution and difficulty level of these components, the goals are to 1) preliminarily discern Year 12 students' development in these components, 2) determine if one component is particularly challenging and 3) compare two items assessing the same skills. This analysis begins with the overall distribution of the three components, as shown in Table 6.3.

Table 6.3

Distribution of Students' Component Levels Across the Hierarchy by Percentage

		Lower group			Upper group		
Component	Year	*L1*	*L2*	*L3*	*L4*	*L5*	*L6*
Text and Context (TnC)	9	2%	8%	38%	48%	4%	0%
	12	0%	0%	25%	58%	17%	0%
Representation (Rep)	9	2%	8%	36%	54%	0%	0%
	12	0%	0%	19%	73%	8%	0%
Statistical-mathematical (SnM)	9	2%	8%	42%	44%	4%	0%
	12	0%	2%	21%	60%	17%	0%

Note. L1 signifies Level 1 etc.

Table 6.3 illustrates the distribution of students across both grades through six levels for each of three components. Four primary observations emerge from these results. Firstly, Year 12 students outperformed Year 9 in all components, evidenced by their higher percentages in the upper group. Secondly, the highest number of Year 9 and Year 12 students was at Level 4, despite the slight difference in the number of Year 9 students between Levels 4 and 3 for statistical-mathematical knowledge. Thirdly, Level 6 remained unachieved by both grades, with Year 9 students were still coded in Levels 1 and 2. Lastly, students in Year 12 found representation to be very easy to decode, given that 81% of students achieved the upper group level.

Those first three findings align with the distribution of students’ SL and four skill levels, whereas the fourth warrants additional exploration. The distribution in Table 6.3 indicates a higher count of Year 12 students in the upper group for representation than the other two components, yet fewer reached Level 5 (critical), underscoring the difficulty in critical graph reading. Many Year 12 students exhibited just appropriate non-critical graph interpretation, challenging the notion that they found representation simpler than other SL components.

Subsequent analysis revealed the frequency distribution for each item's components, useful for discerning component difficulty and examining the unique distribution in interpreting skills. Tables 6.4–6.6 present the frequency distribution of the three components for interpreting, evaluating and decision-making items, excluding communicating skills due to the apparent ease of related items for Year 9 and particularly Year 12 students. The frequency distribution of the three components for communicating skill is available in Appendix L.

Table 6.4

Distribution of Students' Levels in All Components for Interpreting Item Across the Hierarchy by Percentage

Item (Year)	Compo-nent	Lower group			Upper group		
		L1	*L2*	*L3*	*L4*	*L5*	*L6*
The Production Mean (Year 9)	TnC	10%	10%	34%	23%	13%	10%
	Rep	10%	6%	38%	19%	19%	8%
	SnM	8%	17%	44%	13%	8%	10%
The Production Mean (Year 12)	TnC	6%	4%	36%	23%	19%	12%
	Rep	4%	8%	40%	18%	15%	15%
	SnM	4%	10%	42%	13%	14%	17%
The Most Production (Year 9)	TnC	8%	29%	15%	33%	11%	4%
	Rep	8%	29%	15%	36%	6%	6%
	SnM	8%	31%	11%	29%	13%	8%
The Most Production (Year 12)	TnC	8%	13%	6%	44%	19%	10%
	Rep	8%	13%	4%	42%	31%	2%
	SnM	8%	13%	4%	33%	29%	13%

Note. L1 signifies Level 1 etc.

Table 6.4 delves into the distribution specifics for two interpreting items, revealing unique patterns. The Production Mean item saw a notably large proportion of both Year 9 and 12 students at Level 3. In contrast, The Most Production had a substantial number of Year 9 students at Level 2, clarifying their distinct performance in interpreting skills (see Table 6.2). The distribution also confirms that The Production Mean posed more challenges for Year 12 students than The Most Production item, although both assessed interpreting skills within the same context. Further investigation was then required to identify possible factors causing these discrepancies in two interpreting items (The Most Production is considered easier than The Production Mean) for Year 12 students.

Having examined the distribution in Table 6.4, students in both grades generally perceived the three components of each interpreting item to be of similar difficulty levels. For instance, the percentage of Year 9 students in the upper group for the three components of The Most Production item fell between 41% – 60% interval, denoting moderate difficulty. This suggests a close interrelation among those three components, where a challenge in one reflects challenges in others. However, the statistical-mathematical knowledge of The Production Mean item proved to be challenging for Year 9 students, possibly due to procedural understanding of the mean but miscalculations or misinterpretations from the graph.

While Year 9 students struggled equally with both interpreting items, most Year 12 students successfully interpreted The Most Production item. Insights from Table 6.4 include: 1) a shift for Year 9 students from Level 3 in The Production Mean item to Level 2 in The Most Production item, 2) a reduced proportion of Year 12 students in the lower group, particularly Level 3 for The Production Mean item, 3) an increase in Year 12 students at Level 4 across all components of The Most Production item, and 4) a doubling in Year 12 students at Level 5 for representation and statistical-mathematical knowledge in The Most Production. The first three insights underscore the three components' interconnectedness, while the last insight suggests that the question itself impacts Year 12 students' graph interpretation and eventually affects their statistical-mathematical knowledge.

In addition to the interpreting items, Table 6.5 displays the component distribution for evaluating items, indicating The Employees item was moderately challenging for Year 9 students and more challenging than the Mathematics Scores for Year 12 students. Generally, no single component overly influenced students' evaluating skill, except for Year 12 students' ease with representation and the statistical-mathematical component in the Mathematics Scores item. However, the distribution of Year 12 students in The Employees item was different to that of the others. The number of Year 12 students who achieved Level 5 was considerably higher than that for the others, with representation having the lowest among the three components of The Employees item. Year 12 students might struggle making sense of bar graph with discontinued y-axis. Based on the distribution in Level 4 and Level 5 for this item, Year 12 students seemed unable to interpret graphs critically.

Table 6.5

Distribution of Students' Levels in All Components for Evaluating Items Across the Hierarchy by Percentage

		Lower group			Upper group		
Item (Year)	Component	*L1*	*L2*	*L3*	*L4*	*L5*	*L6*
The Employees (Year 9)	TnC	8%	10%	36%	40%	6%	0%
	Rep	10%	10%	28%	46%	6%	0%
	SnM	10%	10%	36%	38%	6%	0%
The Employees (Year 12)	TnC	0%	9%	29%	31%	21%	10%
	Rep	2%	2%	31%	42%	15%	8%
	SnM	2%	6%	36%	25%	25%	6%
Mathematics Scores (Year 9)	TnC	4%	17%	27%	44%	4%	4%
	Rep	8%	15%	35%	34%	4%	4%
	SnM	10%	17%	27%	38%	4%	4%
Mathematics Scores (Year 12)	TnC	2%	2%	17%	65%	6%	8%
	Rep	2%	4%	13%	67%	6%	8%
	SnM	2%	2%	15%	65%	10%	6%

Note. L1 signifies Level 1 etc.

Finally, Table 6.6 outlines the distribution across all components for the decision-making item, indicating a moderate difficulty for Year 9 students but an ease for Year 12 students. It is evident that students from each grade level performed similarly on both The 100-Metre Race and the Which Motorcycle? items. The levels of the three components for a single decision-making item were mostly at the same level of difficulty. However, Year 9 students struggled with the table representation in Which Motorcycle? item, likely due to overlooking the crucial 2.5% tax detail. This oversight led to lower representation knowledge, and even those who noticed the tax struggled to incorporate it into their calculations, revealing a gap in statistical-mathematical understanding. Conversely, Year 12 students easily grasped the context of choosing a motorcycle, simplifying the task for them.

Table 6.6

Distribution of Students' Levels in All Components for the Decision-Making Item Across the Hierarchy by Percentage

	Compo-	Lower group			Upper group		
Item (Year)	nent	*L1*	*L2*	*L3*	*L4*	*L5*	*L6*
The 100-Metre Race (Year 9)	TnC	2%	15%	33%	35%	15%	0%
	Rep	2%	8%	40%	35%	15%	0%
	SnM	2%	10%	42%	35%	9%	2%
The 100-Metre Race (Year 12)	TnC	0%	4%	21%	33%	21%	21%
	Rep	0%	4%	21%	31%	25%	19%
	SnM	0%	6%	21%	35%	23%	15%
Which Motorcycle? (Year 9)	TnC	0%	10%	40%	35%	13%	2%
	Rep	0%	15%	47%	21%	13%	4%
	SnM	0%	21%	37%	27%	11%	4%
Which Motorcycle? (Year 12)	TnC	0%	0%	19%	46%	31%	4%
	Rep	0%	4%	21%	52%	17%	6%
	SnM	0%	2%	29%	36%	27%	6%

Note. L1 signifies Level 1 etc.

6.1.4 Summary

This section concisely presents the findings that address the first research question: 'What levels of SL do Indonesian high school students possess?' Analysis of the data, which categorises students' levels across the six hierarchical levels, yields three primary conclusions.

Firstly, the predominant performance of both Year 9 and Year 12 students was identified at Level 4 (consistent non-critical). This suggests that the highest number of students demonstrated an appropriate but not critical statistical thinking in their responses. Notably, there was a relatively big number of Year 9 students displaying a Level 3 competence (inconsistent), revealing an inconsistent application of statistics in their answers.

Secondly, an extensive portion of students from both grades predominantly spanned Levels 3 and 4 across in almost all skills. However, exceptions were observed in the interpreting skill for Year 9 students and the communicating skill for Year 12 students. Within the interpreting skill, the highest number of the Year 9 students exhibited informal thinking (Level 2). Comparatively, there were no Year 12 students who used limited statistical thinking in communicating their responses, with the highest number of them demonstrating non-critical thinking (Level 4).

Lastly, an analysis between the two grades showed that Year 12 students outperformed their Year 9 counterparts in overall SL, the four skills, each assessment item and three components. This trend suggests a positive progression in students' SL as they advance in grades. However, a detailed statistical examination is essential to validate this observation, a subject that will be explored in the following section.

6.2 Differences in Students' SL

This section reports the findings from Mann-Whitney U tests exploring the differences based on participants' backgrounds. A series of Mann-Whitney U tests were performed due to the ordinal nature of the data (students' level). Further, a Mann-Whitney U test was run following guidance from Laerd Statistics (2015). Participants' grade level and demographic information served as the independent variable, while the dependent variables for all the tests were the SL and the four skills. Section 6.2.1 presents the findings on differences by grade level, complementing the previous descriptive statistics, while Section 6.2.2 presents the findings on gender differences. Sections 6.2.3–6.2.5 present the findings for differences by school type, school status and city of origin, respectively. The summary findings of all statistical analyses are presented in Section 6.2.6 to answer the second research question.

6.2.1 Students' SL by Grade Level

Before presenting the results of Mann-Whitney U tests by grade level, the four assumptions were checked (see Chapter 5, Section 5.5.1.3 for more detail). It was found that one assumption was violated: the distributions of students' levels for Year 9 and Year 12 were (mostly) dissimilar. Therefore, the mean rank was analysed to determine whether there are any statistical differences in students' SL and skill level based on their grade level. Table 6.7 presents the result of the Mann-Whitney U test by grade level.

Table 6.7

Mann-Whitney U Test Results for SL by Grade Level

Variable	Number of students		Mean rank		*U*	*z*	*p*	In favour of
	Y-9	*Y-12*	*Y-9*	*Y-12*				
SL	48	48	40.95	56.05	1,514.5	3.041	.002	Y-12
I	48	48	43	54	1,416	1.983	.047	None
C	48	48	40.72	56.28	1,525.5	3.508	<.001	Y-12
E	48	48	40.31	56.69	1,545	3.065	.002	Y-12
D	48	48	38.34	58.66	1,639.5	3.848	<.001	Y-12

Note. p<.05
SL=Statistical Literacy, I=Interpreting, C=Communicating, E=Evaluating, D=Decision-making; Y-9=Year 9, Y-12=Year 12.

The result shows that the SL level was statistically and significantly different between Year 9 (mean rank = 40.95) and Year 12 (mean rank = 56.05), U = 1,514.5, z = 3.041, p = .002, using an exact sampling distribution for U. When the four SL skills (interpreting, communicating, evaluating and decision-making) were considered separately, using a Bonferroni-adjusted alpha level of .013, the results show statistically significant differences between Year 12 and Year 9 in communicating, evaluating and decision-making. However, there was no significant difference between Year 12 and Year 9 in the interpreting skill. The results of the Mann-Whitney U tests by grade level complemented the Year 9 and 12 students' frequency distribution across the six hierarchical levels (see again Tables 6.1 and

6.2). Particularly, given that both Year 9 and 12 students were spread throughout all six levels only in interpreting, the difference in this skill was determined to be not significant.

Consequently, another Mann-Whitney U test was conducted to explore if there were differences between Year 12 and Year 9 students' levels across the two interpreting items. The same analyses were also performed with the remaining six items, as presented in Table 6.8.

Table 6.8

Mann-Whitney U Test Results for Eight Items by Grade Level

Variable	Number of students		Mean rank		*U*	*z*	*p*	In favour of
	Y-9	*Y-12*	*Y-9*	*Y-12*				
The Production Mean (I)	48	48	46.26	50.74	1,259.5	0.818	.414	None
The Most Production (I)	48	48	41.41	55.59	1,492.5	2.579	.010	Y-12
YouTube Viewers (C)	48	48	39.88	57.13	1,566	3.553	<.001	Y-12
Domestic Waste (C)	48	48	44.96	52.04	1,322	1.494	.135	None
The Employees (E)	48	48	41.41	55.59	1,492.5	2.599	.009	Y-12
Mathematics Scores (E)	48	48	39.86	57.14	1,566.5	3.314	<.001	Y-12
The 100-Metre Race (D)	48	48	39.04	57.96	1,606	3.454	<.001	Y-12
Which Motorcycle? (D)	48	48	39.52	57.48	1,583	3.289	.001	Y-12

Note. p<.05
I=Interpreting, C=Communicating, E=Evaluating, D=Decision-making; Y-9=Year 9, Y-12=Year 12.

The result shows there was no significant difference by grade in The Production Mean item and a significant difference by grade in The Most Production item. This explains the frequency distribution of these two items (see again Table 6.3). Particularly, although students in both grade levels were spread over the six levels, the distributions were similar and proportional for The Production Mean item but not proportional for The Most Production item. Further, the result shows that the two communicating items have different results, despite there was a statistically significant difference in students' overall communicating skill. The result also shows that there was significant difference by grade in the YouTube Viewers item in favour of the Year 12 students and no significant difference by grade in the Domestic Waste item.

6.2.2 Students' SL by Gender

The third Mann-Whitney U test was conducted to explore if there were differences in participants' SL and skill levels between boys and girls, with the result presented in Table 6.9.

Table 6.9

Mann-Whitney U Test Results for SL by Gender

Variable	Number of students		Mean rank		*U*	*z*	*p*	In favour of
	Boys	*Girls*	*Boys*	*Girls*				
SL	48	48	48.23	48.77	1,165	.109	.913	None
I	48	48	53.39	43.61	917.5	–1.761	.078	None
C	48	48	50.25	46.75	1,068	–.789	.430	None
E	48	48	46.85	50.15	1,231	.616	.538	None
D	48	48	51.70	45.30	998.5	–1.212	.226	None

Note. p<.05
SL=Statistical Literacy, I=Interpreting, C=Communicating, E=Evaluating, D=Decision-making.

The result shows that SL level was not statistically significantly different between boys (mean rank = 48.23) and girls (mean rank = 48.77), *U* = 1,165, *z* = .109, *p* = .913, using an exact sampling distribution for *U*. When the four SL skills (interpreting, communicating, evaluating and decision-making) were considered separately using a Bonferroni-adjusted alpha level of .013, the results again show that the four skills' levels (interpreting, communicating, evaluating and decision-making) were not statistically significantly different between boys and girls.

6.2.3 Students' SL by School Type

The fourth Mann-Whitney U test was conducted to explore if there were differences in participants' SL and skill levels between schools under MoRA and schools under MoEC-RT. The result is presented in Table 6.10.

Table 6.10

Mann-Whitney U Test Results for SL by School Type

Variable	Number of students		Mean rank		*U*	*z*	*p*	In favour of
	MoRA	*MoEC-RT*	*MoRA*	*MoEC-RT*				
SL	48	48	44.41	52.59	1,348.5	1.649	.099	None
I	48	48	44.90	52.10	1,325	1.299	.194	None
C	48	48	46.92	50.08	1,228	.714	.475	None
E	48	48	48.26	48.74	1,163.5	.090	.929	None
D	48	48	45.18	51.82	1,311.5	1.259	.208	None

Note. p<.05
SL=Statistical Literacy, I=Interpreting, C=Communicating, E=Evaluating, D=Decision-making; MoRA=Ministry of Religious Affairs, MoEC-RT=Ministry of Education, Culture-Research and Technology.

The result shows that the SL level was not statistically significantly different between schools under MoRA (mean rank = 44.41) and schools under MoEC-RT (mean rank = 52.59), *U* = 1,348.5, *z* = 1.649, *p* = .099, using an exact sampling distribution for *U*. When the four SL skills (interpreting, communicating, evaluating and decision-making) were considered separately using a Bonferroni-adjusted alpha level of .013, the results show no statistically significant difference across all skills' levels (interpreting, communicating, evaluating and decision-making) between students from schools under MoRA and students from schools under MoEC-RT.

6.2.4 Students' SL by School Status

The fifth Mann-Whitney U test was conducted to explore if there were differences in participants' SL and skill levels between public schools and private schools. The result is presented in Table 6.11.

Table 6.11

Mann-Whitney U Test Results for SL by School Status

Variable	Number of students		Mean rank		*U*	*z*	*p*	In favour of
	Public	*Private*	*Public*	*Private*				
SL	48	48	49.05	47.95	1,125.5	–.222	.824	None
I	48	48	46.45	50.55	1,250.5	0.740	.459	None
C	48	48	48.34	48.66	1,159.5	0.070	.944	None
E	48	48	48.88	48.13	1,134	–.140	.888	None
D	48	48	50.40	46.60	1,061	–0.718	.473	None

Note. p<.05
SL=Statistical Literacy, I=Interpreting, C=Communicating, E=Evaluating, D=Decision-making.

The result shows that SL level was not statistically significantly different between public schools (mean rank = 49.05) and private schools (mean rank = 47.95), *U* = 1,125.5, *z* = –.222, *p* = .824, using an exact sampling distribution for *U*. When the four SL skills (interpreting, communicating, evaluating and decision-making) were considered separately using a Bonferroni-adjusted alpha level of .013, the results again show that the four skills' levels (interpreting, communicating, evaluating and decision-making) were not statistically significantly different between public schools and private schools.

6.2.5 Students' SL by City of Origin

The last Mann-Whitney U test was conducted to explore if there were differences in participants' SL and skill levels between students in a non-metropolitan city (Jombang) and students in a metropolitan city (Surabaya). The result is presented in Table 6.12.

Table 6.12

Mann-Whitney U Test Results for SL by City of Origin

Variable	Number of students		Mean rank		*U*	*Z*	*p*	In favour of
	Jbg	*Sby*	*Jbg*	*Sby*				
SL	48	48	48.95	48.05	1,130.5	–0.180	.857	None
I	48	48	47.24	49.76	1,212.5	0.454	.650	None
C	48	48	51.15	45.85	1,025	–1.193	.233	None
E	48	48	50.88	46.13	1,038	–0.889	.374	None
D	48	48	49.40	47.60	1,109	–.339	.734	None

Note. p<.05
SL=Statistical Literacy, I=Interpreting, C=Communicating, E=Evaluating, D=Decision-making;
Jbg = Jombang (non-metropolitan city), Sby = Surabaya (metropolitan city).

The result shows that SL level was not statistically significantly different between students from a non-metropolitan city (mean rank = 48.95) and students from a metropolitan city (mean rank = 48.05), *U* = 1,130.5, *z* = –.180, *p* = .8571, using an exact sampling distribution for *U*. When the four SL skills (interpreting, communicating, evaluating and decision-making) were considered separately using a Bonferroni-adjusted alpha level of .013, the results again show there were no statistically significant differences in all skills' levels (interpreting, communicating, evaluating and decision-making) between students from a non-metropolitan city and students from a metropolitan city.

6.2.6 Summary

This section addresses the second research question: 'Are there any significant differences in Indonesian high school students' SL based on their backgrounds (i.e., grade level, gender, school type, school status or city of origin)?' Several conclusions can be drawn from the Mann-Whitney U tests results. The analyses revealed statistically significant differences in students' SL levels based on grade level, specifically favouring Year 12 students. No statistically significant differences were found in SL levels concerning the other variables (gender, school type, school status and city of origin).

Further examination of the four response skills indicated that statistically significant differences were observed solely in relation to grade levels. Specifically, Year 12 students scored statistically higher in every SL skill, except for interpretating. This necessitated an additional analysis to understand the absence of statistically significant differences in the interpretation skill levels between Year 12 and Year 9 students. A likely explanation for this is the students' challenges in interpreting data from line graphs, especially those with embedded context.

6.3 Discussion

This section presents three discussion topics that emerged from the quantitative findings. Section 6.3.1 discusses the development in Indonesian students’ SL level across grade levels as well as the differences of their SL levels based on the other demographics backgrounds. Section 6.3.2 addresses the concern on the students’ future participation in an information-driven society, given their absence in the critical mathematical level—the highest level. Section 6.3.3 discusses the applicability of the proposed SL assessment framework for teachers to both investigate students’ SL levels and facilitate students’ learning.

6.3.1 Development and Differences in Indonesian Students’ SL Levels

This study employed a cross-sectional design to explore the influence of grade level—Year 9 and Year 12—on Indonesian students’ SL level and their level in the four SL skills: interpreting, communicating, evaluating and decision-making. Demographic factors were also assessed. The key findings indicated 1) Year 12 students’ SL and skill level were statistically higher than that of Year 9 students, but not in interpreting skill, 2) no significant difference was found between boy and girl students in their SL and four skills’ levels, and 3) no significant differences were found in students’ SL and four skills’ level by the other backgrounds (school type, school status and city of origin). The discussions on the abovementioned topics are presented below.

First, contrary to prior research (Aoyama & Stephens, 2003; Callingham & Watson, 2017; Yolcu, 2014), which suggested that non-adjacent grade levels would naturally correspond with improved SL levels due to cognitive development and external data interactions, this was not substantiated for interpretation skills. Both Year 9 and Year 12 students demonstrated challenges in this particular skill. The data revealed limited statistical thinking among these students, signifying that interpreting data-based information remains a universal difficulty.

A deeper inquiry into this anomaly suggests that the complexity of the task, particularly the line graph utilised, might be a confounding factor. Line graphs are known to pose interpretative challenges (Ali & Peebles, 2013; Bursal & Yetiş, 2020; Patahuddin & Lowrie, 2019; Peebles & Ali, 2015), a sentiment echoed by this study. This is consistent with long-standing deficiencies in the interpretation of line graphs among Indonesian students (TIMSS & PIRLS, 2011). Accordingly, pedagogical interventions targeting the critical interpretation of complex line graphs are strongly recommended.

Second, this study provided a current perspective on the gender variable of Indonesian high school students' SL. The current study findings corroborate large-scale studies, such as PISA 2003 and 2012, which report no gender disparities in Indonesian students' SL levels (OECD, 2004, 2014). However, the absence of gender differences is not a cause for celebration. Both genders performed poorly in problems involving statistics, contrasting sharply with ASEAN counterparts like Singapore and Vietnam, where both genders scored high (OECD, 2014).

The research was conducted post-implementation of Indonesia's Curriculum 2013 (K13), which had undergone significant modifications to improve students' performances in the international test and make the curriculum more relevant to their lives. Despite these changes, the study suggests that improvements may have fallen short of expectations. Moreover, the participants of this study were selected from East Java province, whose provincial average score in the UN 2019 test was slightly above the national average score for mathematics (Pusat Penilaian Pendidikan Kemdikbud, 2023). Therefore, the results may not be representative of underprivileged areas, and other provinces might yield lower performance levels.

Third, unexpectedly, school type, school status or city of origin showed no statistical influence on SL levels. This contradicts with prevailing Indonesian perceptions that private and MoRA-affiliated schools underperform relative to public and MoEC-RT-affiliated schools (Bedi & Garg, 2000; Muttaqin et al., 2020; Newhouse & Beegle, 2006). One rationale may be the study's geographical focus was on East Java, a region with above-average educational metrics (Azzizah, 2015; Pusat Penilaian Pendidikan Kemdikbud, 2023), suggesting the need for broader geographic sampling in future studies.

6.3.2 Concern on Students' Future Participation in Information-Driven Society

This study revealed a noteworthy progression in performance from Year 9 to Year 12 students, with the highest proportion of both grades showed consistent but non-critical thinking (Level 4). This finding aligns with previous research such as Callingham and Watson (2017). Specifically, they confirmed there was a trend in the development of students' SL across grade levels (from Year 5 to Year 10), but most Year 10 students still performed in Level 4. While the prior and this study indicated the advancement of students' SL across grade levels, it raised concerns about students' limitations in critical thinking.

Such observations raise questions about the preparedness of Year 12 students to engage effectively in an information-driven society post-education. The concern emanates not merely from the median achievement level, which approaches Level 4, but also from the complete absence of students at the critical mathematical level (Level 6). The lack of Year 9 students at this level aligns with their historical underperformance in the PISA assessments on uncertainty and data (OECD, 2004, 2014, 2023). However, the absence of Year 12 students at Level 6 is alarming for two reasons: 1) these students, at the end of formal schooling, are expected to possess the ability for critical interpretation of data-based information, as stipulated by Gal (2002) and Watson & Callingham (2020), and 2) their SL levels set the standard for future societal engagement, as they represent the absent of educational outcomes anticipated by Curriculum 2013 (K13)—implemented a decade prior to this study.

Given the current state of SL among Indonesian high school students, as evidenced by both Year 9 and Year 12 cohorts, urgent and substantial interventions are requisite. These interventions must involve all educational stakeholders to foster an environment conducive to the development of SL. Challenges specific to SL—as will be detailed in Chapter 7—should be adequately addressed. The ultimate aim is to enable each student to exhibit the four data consumption skills, preparing them for informed participation in society.

6.3.3 Applicability of the Proposed SL Assessment Framework for Teachers

The SL assessment framework proposed in this study offers a new construct to assess students' SL from the data consumer perspective. This framework was developed and then applied to assess four skills students need as consumers of data-based information: interpreting, communicating and evaluating statistical information and using statistical information in making informed decisions (Budgett & Rose, 2017; Franklin et al., 2005; Gal, 2002; Guler et al., 2016; Wallman, 1993). This framework supports the objectives of national curriculum regarding the application of students' statistical knowledge in society (Kemdikbud, 2012). This framework also complements the mathematics textbooks, as the four SL skills correspond to the competencies that Indonesian high school students are expected to attain (see Chapter 3). Data-based problems assessing these skills have been found across high schools' mathematics textbooks (e.g., As'ari et al., 2016, 2017a, 2018; Sinaga et al., 2014a, 2014b; Subchan et al., 2015). These problems were open ended and designed to help teachers organising their classrooms through 5L activities: let's observe, let's ask, let's find information, let's reason and let's share. It is consequently expected that students will have more opportunities to engage with these skill-based problems and become increasingly statistically literate.

However, although some skill-based problems were identified in textbooks, most of students could not provide critical responses as reflected in this study's findings—being in Level 4 (consistent non-critical) or below. Theoretically, mathematics textbooks show the link between the intended and implemented curriculum (Valverde et al., 2002; Weiland, 2019) and strongly influence mathematics teaching and learning (Büscher, 2022b; Landtblom, 2018; Ponte & Marques, 2011). In addition, textbooks are the primary resources for teachers and students (Landtblom, 2018; Reys et al., 2004). Having provided with mathematics textbooks that support students' SL (see Chapter 3), it was then the mathematics teacher who were expected to play essential role in teaching students with SL. In addition, they were expected to be able to conduct an assessment to inform students' achievement, facilitate student learning (Sabbag et al., 2018) and reflect on students' cognitive thought when responding to data-based problem. Although assessing students' ability to think and reason statistically could be challenging (Woodard et al., 2020), their ability to think and reason statistically is of the intended learning outcomes.

In response to this need, the process of and results obtained from this study ensure the applicability of this new SL assessment for education stakeholders, particularly for teachers. This study clearly described how a test was conducted to investigate students' SL levels using skill-based items. Further, the process of determining students' SL levels from their written responses were clearly and procedurally explained. Mathematics teachers could replicate this assessment process with their own students in the class. They could select statistical problems from the textbooks, newspaper or design their own problems that can be used to assess students' particular skill in relation to the three components. However, these selected skill-based problems should provide opportunities for students to respond across the six hierarchical levels—as those used in this study. Thus, the level descriptors in the form of conjectured answers regarding the three SL components need to be developed.

Alternatively, mathematics teachers could also use this framework to facilitate students' learning. Teachers can use 5L classroom activities to target individuals' learning progress and align instruction to the specific skill (e.g., communicating relevant information and evaluating statistical claim). In this case, the problems do not necessarily to cover the six hierarchical levels. Teachers only need to identify the highest level those problems can cover and use it to guide the statistics instruction. In a particular case, the problems also do not necessarily to involve the three SL components. This considers Gal and Geiger's (2022) views that not all SL items require the interpretation of a graphic. The most important thing is to ensure that every student learns how to critically respond to data-based information.

6.4 Chapter Summary

This chapter elucidates the findings for the quantitative component of this study, and general conclusions can be drawn from the results. Regarding the first research question, the highest number of Year 9 and Year 12 students performed in Level 4 (consistent non-critical), which signifies non-critical but appropriate application of statistical thinking. Notably, Year 12 students outperformed their Year 9 counterparts in overall SL levels as well as across four distinct skills, suggesting a development of students' SL across grades. Pertaining to the second research question, statistical analyses verified significant disparities in SL and skill level based on grade, with the exception of the interpreting skill. Intriguingly, no such differences were observed when considering other demographic factors, such as gender, school type, school status and city of origin.

Based on those findings, a discussion was held regarding the students' development and readiness for social participation as well as the potential applications of the SL assessment framework. While Year 12 students showed improvement in their SL compared to Year 9 students, their limitation in critical thinking may have fallen short of expectation. This is not what was expected, since the curriculum that was implemented ten years ago was meant to enhance their critical thinking. As such, their engagement in a society is dubious. In response to this issue, education stakeholders—especially the mathematics teacher—may decide to monitor students' SL and support their learning using the SL assessment framework proposed in this study.

It is crucial to consider the performance of students classified within the lower group (Levels 1–3) irrespective of their grade. These students appear to face challenges in comprehending the three SL components and may struggle with these components. The subsequent chapter will delve into the qualitative aspects of the study, concentrating on these challenges while also elucidating the appropriate and critical understandings of students in the upper group (Levels 4–6).

Chapter 7: Students' Challenges and Understandings in Responding to Statistical Information

This chapter elucidates the findings derived from the analysis of written responses and interview data of 24 students, with varied SL levels. This chapter addresses the study's third and fourth research questions: 'How do the challenges students encounter in comprehending the three components of SL affect their abilities to respond to statistical information?' and 'How do students' understandings of the three components of SL influence their abilities to respond to statistical information?' The SL components in focus are textual and contextual understanding, representation and statistical-mathematical knowledge.

To maintain coherence with the quantitative results, the 24 students are categorised into two groups: the lower group (Levels 1–3) and the upper group (Levels 4–6). The demarcation between these groups provides an insightful lens through which students' different responses can be assessed. Qualitative analysis of students' written responses and interview data in the lower group would allow identifying the challenges they encountered in making sense of the three components of SL items. In contrast, analysis of students' responses in the upper group would show what enabled them to solve the items, particularly regarding the three components of SL items.

The following sections are primarily intended to reveal students' challenges compared to their understandings of statistical information. However, it is important to begin with the context of the 24 students, provided in Section 7.1. This section provides fundamental findings and information on the levels the 24 students achieved in terms of overall SL, the four skills and the eight items. Such information is crucial to validate the sample's representativeness, especially considering the post-test selection of participants, when their SL levels had not yet been determined. In addition, it is important to ensure that there were students in both the lower and upper groups for which their written works and interviews at a certain level could be exemplified. Section 7.2 presents the general challenges students encountered in comprehending the three components, while Section 7.3 presents the students' overall understandings in the same regard.

The chapter also posits that the challenges or understandings in SL might share commonalities in relation to the skills. Both interpreting and communicating necessitate students to articulate their critical comprehension of the SL components, albeit for different audiences: for themselves (interpreting) and others (communicating). Similarly, the skills of evaluating and decision-making require substantiating evidence drawn from the SL components in order to contest claims (evaluating) or to justify their choices (decision-making). As such, Sections 7.2 and 7.3 have subsections for interpreting and communicating, as well as evaluating and decision-making. In each section, empirical evidence in the form of student-written work and interviews are used to illustrate the challenges or understandings exhibited by the students in making sense of the three knowledge components. Section 7.4 discusses the key findings, and Section 7.5 offers a summary of the entire chapter.

7.1 Context of the Interviewed Students

The 24 students interviewed were selected from the total study participants (n = 96) to represent the variety of students' SL levels. Unlike the quantitative findings, where the students were separated by grade level, the qualitative findings combined students to represent the hierarchical level, as either Year 9 or Year 12 students exposed similar thought patterns when at the same level. Table 7.1 reveals that these 24 students were representative in two ways: a high percentage of students fell within Levels 4 and 3; and their distribution across SL levels closely mirrored the broader sample of 96 students (see again Table 6.1 in Chapter 6), affirming the validity of the findings.

Table 7.1

Distribution of the Interviewed Students Across the Hierarchical Levels

Level	N	%
L2 (informal)	1	4%
L3 (inconsistent)	5	21%
L4 (consistent non-critical)	13	54%
L5 (critical)	5	21%

Note. L2 signifies Level 2 etc.

As students' responses are presented in association with the four skills, it is important to reveal their levels in each skill and each of the two items for the respective skills. Table 7.2 delineates the distribution of levels achieved by the 24 students across the four SL skills. This data stands in relation to Table 6.2 from Chapter 6, which documents the same for the full sample of 96 students. The analysis between the two tables validates that the selected 24 students were representative, closely approximating the distribution of skill levels among the larger cohort. Nevertheless, it should be noted that no students from the interview cohort represented Level 2 'communicating' and Level 1, Level 5 and Level 6 'evaluating', despite a few students across the larger cohort performing at those levels. This limitation is not considered to be a significant issue, given the representativeness of students across other skills' levels. Moreover, the data in Table 7.3, which presents the distribution of levels achieved by the 24 students in all eight items, filled this gap. The distribution of students in this table ensures that the qualitative data analysed are sufficiently varied to represent data from the larger cohort, particularly to represent Level 1, Level 5 and Level 6 'evaluating'.

Table 7.2

Number of Interviewed Students Across the Hierarchy for the Four Skills

	L1	L2	L3	L4	L5	L6
Interpreting (I)	1	3	4	5	7	4
Communicating (C)	–	0	5	16	3	–
Evaluating (E)	0	4	9	11	0	0
Decision-making (D)	–	2	8	10	3	1

Note. L1 signifies Level 1 etc, – signifies the absence of students from both the larger and interviewed cohort in that level, and 0 signifies the absence of students from the interviewed cohort in that level despite the presence of students from the larger cohort in that level.

Table 7.3

Number of Interviewed Students Across the Hierarchy for the Eight Items

	L1	L2	L3	L4	L5	L6
The Production Mean (I)	1	0	9	5	3	6
The Most Production (I)	2	2	1	4	8	7
YouTube Viewers (C)	1	0	3	12	8	–
Domestic Waste (C)	–	0	5	15	4	–
The Employees (E)	0	3	9	7	4	1
Mathematics Scores (E)	3	3	4	11	3	0
The 100-Metre Race (D)	–	5	4	5	6	4
Which Motorcycle (D)	–	3	9	7	4	1

Note. L1 signifies Level 1 etc, – signifies the absence of students from both the larger and interviewed cohort in that level, and 0 signifies the absence of students from the interviewed cohort in that level despite the presence of students from the larger cohort in that level.

Finally, the students' levels on each item are used to present the students' responses. This choice considered that the skill level was the median of their levels in the three components of the two items assessing such skill. Further, although most students achieved the same levels or one-level difference in each of the two items assessing the same skill, some students achieved levels with a two- or three-level difference. Using skill level would not be practical if the students performed in the lower group for the first item and in the upper group for the second item. Hence, using the level that the students achieved on each item is more practical for making their written works easier to use as relevant examples.

7.2 Students' Challenges in Making Sense of the Three Components

Making sense of the three components across the eight items proved to be challenging for the students in the lower group (Levels 1–3), inhibiting their capacity to adequately respond to statistical information. To delve deeper into these challenges, their responses were analysed. Tables 7.4–7.6 outline these challenges. They were generalised from the students' responses across the eight items that varied in context, representation, statistical-mathematical knowledge and the skills assessed.

Table 7.4

Lower Group Students' Challenges in Making Sense of the Text and Context

Challenges with text and context	Example	Item
Text		
Student inappropriately uses some general terms involved in the text. In addition, the student fails to address instruction being asked in the question, which relates to the four skills.	Student ignores the word 'constant', which refers to the rate of shoe production across the times.	The Production Mean
	Student fails to understand the idea of 'increase', 'time interval' and 'most production' within the context of shoe production.	The Most Production
	Student believes 'huge increase' means 'increase' in the context of an increase in the number of employees.	The Employees
	Student ignores the provided mean of mathematics test score for the two classes being compared yet performs calculation to find a new mean to compare.	Mathematics Scores
	Student misunderstands the instruction in the question 'write summary information' with data listing.	Domestic Waste and YouTube Viewers

Table 7.4 (continued)

Challenges with text and context	Example	Item
Context		
Student fails using the real-life context involved in the problems.	Student shows a lack of understanding of the idea of an increase in shoe production across times.	The Production Mean and The Most Production
	Student misunderstands the context of running competition by thinking that a runner with the highest time is the best.	The 100-Metre Race
	Student categorises burning domestic waste as a proper method because watching others do it.	Domestic Waste
	Student thinks the absence of two bands in January is caused by no one watching their songs. In fact, no singles were released.	YouTube Viewers
	Student thinks the distance travelled by a motorcycle is how far a motorcycle can run with a full tank.	Which Motorcycle?

Table 7.4 suggests that students might find it challenging to understand a single term or the overall context. Students with misconceptions of terminology had personal/individual misunderstandings of the problem. For instance, in The Most Production item, students equated the ‘most production’ as ‘the highest number’ of shoes produced. This misunderstanding led them to search for the highest number on the y-axis of the line graph. In addition, students encountered challenges comprehending the context in which the data emerged. For example, in The 100-Metre Race item, students failed to understand the rules of the running competition that differentiate it from other games (i.e., in the running competition, the quickest are those with the lowest time rather than those with the highest time).

These challenges in understanding the text and context become distractors when used to make sense of the represented data, due to the context being embedded in the graph. Additionally, graphical and tabular conventions pose further challenges, as outlined in Table 7.5. For example, if a student selected a runner with the highest time (in The 100-Metre Race item) or did not understand data increments in the line graph across time (in The Production Mean item), then the student's challenges were influenced by their misunderstanding of the context. Conversely, if the students overlooked the y-axis discontinuity (The Employees item), then their challenges were rooted in graphical convention.

Table 7.5

Lower Group Students' Challenges in Making Sense of Representation

Challenges with representation	Example	Item
The graph/table conventions		
Student informally or inconsistently makes sense of: the title of the graph/table to draw on the general information being presented, the label in axes/row-column to understand what kind of data being compared, the scale in axes/row-column to understand the units, the legends (if there are more than one data) to understand what each data presented and any additional information.	Student ignores the legend (dotted versus solid line, the four bands, class A vs B, etc.)	The Production Mean, The Most Production, YouTube Viewers and Mathematics Scores
	Student ignores the label for the y-axis showing the number of YouTube viewers in thousands and the label for the y-axis showing the percentage for each action of domestic waste management.	YouTube Viewers and Domestic Waste
	Student ignores the discontinuity in the y-axis and uses the difference of bars that almost double as justification.	The Employees
	Student misunderstands the rows and columns, and the data presented for three runners.	The 100-Metre Race
	Student ignores the price of four motorcycles presented in a table, which is in millions of Rupiah.	Which Motorcycle?

Table 7.5 (continued)

Challenges with representation	Example	Item
What does the bar/line/table represent related to context?	Student fails to identify that the data presented in the line graph is discrete and cumulative.	The Production Mean
Student informally or inconsistently: makes sense of data presented; identify how many? How do they and how many they differ? What do the most and the least data mean? Where do they change? And why do they change?	Student fails to notice data points, data increments, constant increase and slope.	The Most Production
	Student thinks the data for three runners are the scores instead of the time in seconds.	The 100-Metre Race
	Student thinks if no bars represent the students' mathematics scores, it means some students did not follow the mathematics test.	Mathematics Scores
	Student thinks burning domestic waste is the most proper method just because its bar shows the highest percentage.	Domestic Waste

Finally, students' challenges in the first two components related to their statistical and mathematical knowledge. Table 7.6 separately shows these challenges across the various items to pinpoint whether these challenges lie in their statistical knowledge, mathematical knowledge or both. It was found that although students demonstrated proficiency in numerical operations, they often misapplied or misinterpreted key elements of statistical knowledge across most of the items. For instance, while they may know a statistical term, they frequently interpret it out of context or with incorrect data reading, resulting in calculations that are procedurally accurate but contextually incorrect.

Table 7.6

Lower Group Students' Challenges in Making Sense of Statistical-Mathematical Knowledge

Challenges with statistical and mathematical knowledge	Example	Item
Statistical knowledge		
Student does not understand statistical ideas that they can use in their responses. When they know some statistical ideas (e.g., mean, mode, patterns and trend), they informally or inconsistently use them in their responses.	Student merely understands the procedure to find a mean using add-divide formula.	The Production Mean
	Student does not understand statistical meaning of the sharpest slope.	The Most Production
	Student does not understand the idea of an outlier that has an impact on the mean of mathematics' test score for the class.	Mathematics Scores
	Student focuses mostly on the most and least data to compare domestic waste.	Domestic Waste
Mathematical knowledge		
Student can perform simple number operation. However, when the numbers being operated are complex (such as involving decimals, percentage, inequality and large numbers), some errors may be found due to their lack of number sense. The calculation they performed was used for inconsistent comparison and grouping.	Despite the calculated mean for shoe production being contextually incorrect, student's calculation is procedurally correct.	The Production Mean
	Student cannot estimate the data when the scale in the axis is big.	YouTube Viewers
	Student fails to calculate the mean for the three runners as the time is in decimal.	The 100-Metre Race
	Student fails to calculate the tax, which is in decimal and percentage, and add it to the price.	Which Motorcycle?

In summary, while the challenges students faced in understanding the three components can be individually identified and distinguished, they are interrelated. This is due to the context being embedded in the representation, and both influencing the statistical-mathematical knowledge. In essence, challenges in one component can affect challenges in the others. Subsequent sections will delve into these challenges in greater detail, drawing upon students' written responses and interviews. These sections will particularly focus on Level 3 (inconsistent) challenges—in relation to the four skills being assessed—given that this level contains the highest percentage of students in the lower group.

7.2.1 Students' Challenges in Interpreting and Communicating Data-Based Information

Students faced challenges in identifying statistical ideas within the presented data and under the provided context, particularly when interpreting quantitative information for themselves. Although both interpreting items—The Production Mean and The Most Production—utilised the same context and line graph; the quantitative findings revealed that the former was moderately difficult for Year 12 students. However, both items posed moderate difficulty for Year 9 students. According to data previously presented in Table 7.3, 10 students were at Level 3 across those two items, while five were at Levels 1 and 2.

These 10 students at Level 3 interpreting struggled to correlate the context with the data represented in the line graph (see work sample from Dani described in the following two pages). By contrast, these students demonstrated a comprehension of statistical measures: the *mean* and its procedure in The Production Mean item and the *mode* and the proportional comparison in The Most Production item. Among those 10 students, nine students were at Level 3 in The Production Mean item, and one was at Level 3 for The Most Production item. These 10 students encountered challenges in determining the total number of shoes produced during the specified times, and two of them were unable to determine both the total shoes and total production time.

These Level 3 students' challenges in making sense of and utilising the data in the representation were influenced by their contextual and graphical misunderstanding. They misinterpreted the data with the context embedded and showed limited knowledge of the graph's convention, which is a line graph. They tended to interpret the data presented in the line graph as if those data were in the bar graph and did not understand the role of the line graph in terms of its relationship with time. For instance, in The Production Mean item, these students appeared to calculate the mean using the arithmetic formula (add-divide formula), indicating a procedural understanding. However, students' lack of contextual meaning of an increase caused them to incorrectly interpret the data increments in the line graph. The gap in their contextual understanding led to flawed statistical thinking and, consequently, incorrect results (see Table 7.7).

Table 7.7

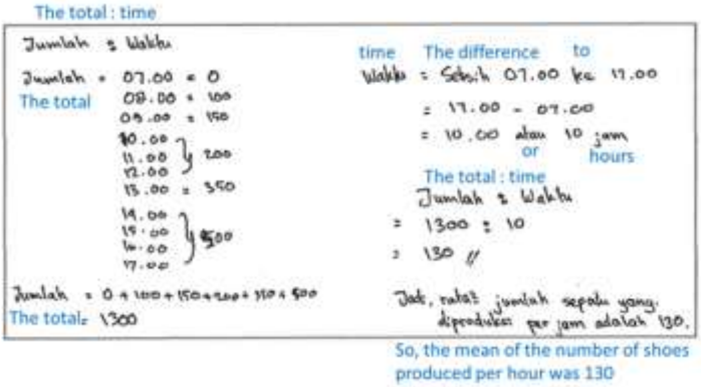
Level 3 Students' Challenges in the Three Components of the Interpreting Skill

	Text and context	Representation	Statistical-mathematical knowledge
The Production Mean	Student shows a lack of understanding of the context of shoe production across times. Student focuses merely on the solid line and ignores the dotted line. Student ignores the word 'constant' and does not understand the idea of 'increase'.	Student could not notice the idea of an increase in a line graph. Student fails to notice that the data is discrete and cumulative. Student fails to notice data points, data increments, and constant increase.	Student does not recognise the resultant mean is incorrect, despite performing a correct number operation, especially addition and division.
The Most Production	Student ignores the word 'constant' and does not understand the idea of 'increase', 'time interval' and 'most production'.	Student fails to notice data points, data increments, constant increase and slope.	Student does not understand the most production within the shortest time. Student does not understand the statistical meaning of the sharpest slope.

Figure 7.1 exemplifies the responses of Dani, a Level 3 student, who found challenges in interpreting a mean from a line graph for The Production Mean item. Dani struggled with understanding the data increments, which affected his calculation of the total number of shoes represented by the solid line. This difficulty was partly due to his inability to comprehend the context embedded in the line graph, as well as the graph's conventions. While Dani did identify the number of shoes based on data points—indicating some understanding of the context and graph conventions—he misunderstood the incremental value of each data point. As a result, he arrived at an incorrect total number of shoes and an incorrect mean, '130 = 1300:10'.

Figure 7.1

Dani's Approach to Solve The Production Mean Item



The total : time
Jumlah : Waktu

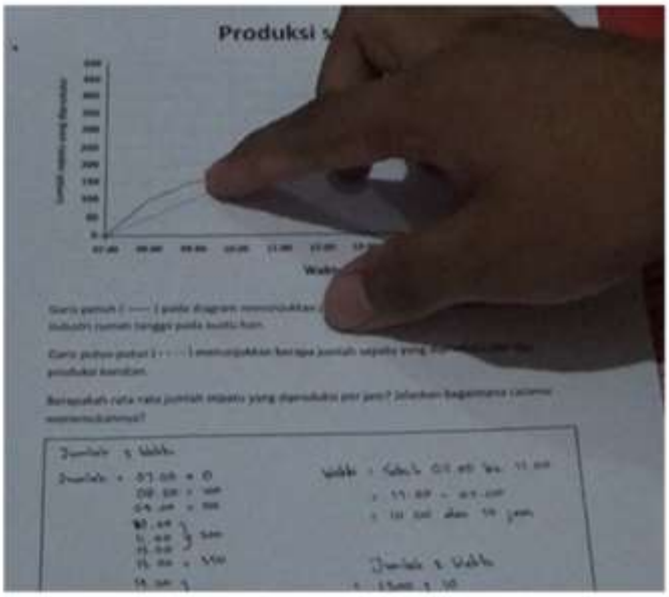
The total
Jumlah = 07.00 = 0
08.00 = 100
09.00 = 150
10.00 11.00 12.00 } 200
13.00 = 350
14.00 15.00 16.00 17.00 } 500

The total
Jumlah = 0 + 100 + 150 + 200 + 350 + 500
= 1300

time The difference to
Waktu = Selisih 07.00 ke 17.00
= 17.00 - 07.00
= 10.00 atau or 10 jam hours

The total : time
Jumlah : Waktu
= 1300 : 10
= 130 //

Jadi, rata² jumlah sepatu yang diproduksi per jam adalah 130.
So, the mean of the number of shoes produced per hour was 130



From the interview, it was evident that Dani (D) based his calculations on the data points represented by the solid line, treating them as non-cumulative. He failed to discern the relationship between the x and y axes in the line graph and the cumulative increase in shoe production. His response, ‘At 09.00, 150 shoes,’ revealed his misunderstanding of the concept of increase represented in the line graph (see again Figure 7.1). He failed to recognise that 150 shoes were produced between 07.00 and 09.00. Additionally, his interpretation of the time interval was flawed; his answer ‘At 10.00 to 12.00, 200 shoes’ should have been ‘From 09.00 to 12.00, 50 shoes.’

I: *Can you show me how many shoes at a certain time?*

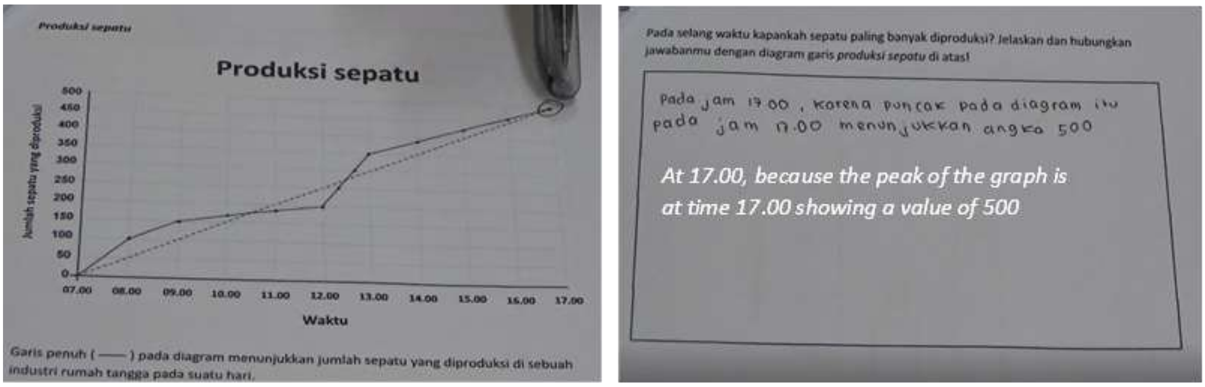
D: *At 08.00, 100 shoes. At 09.00, 150 shoes. At 10.00 to 12.00, 200 shoes. At 13.00, 350 shoes. At 14.00 to 17.00, 500 shoes. And then all together are added, and the total is 1,300 shoes.*

From the above examples of student’s challenges, it can be concluded that students at Level 3 showed an initial understanding of statistics but struggled with data reading. Their statistical and mathematical understanding was inconsistent. For instance, they knew the formula and procedure for calculating the *mean* but inputted incorrect data. Likewise, the other Level 3 students demonstrated similar challenges as Dani, such as being unable to recognise an increment by hours or to determine the total hours. Consequently, the mean was incorrectly calculated from $\frac{2845}{10}, \frac{2925}{10}, \frac{3230}{10}, \frac{2930}{10}, or\ \frac{1200}{5}$. This further indicated that they started to understand the statistical idea but misinterpreted the data presented in the graphs. Such challenges are characteristic of Level 3 students; thus, if they missed them completely in their responses, they would be in Level 2 or Level 1.

Further analysis revealed that Level 1 and Level 2 students could only interpret the values provided in the line graph without any contextual or statistical-mathematical understanding. They completely misinterpreted the context of the problems, failed to read data from the graph, failed to relate data to the context, and lacked statistical and mathematical concepts such as mean, mode and trend. Students' use of colloquial understanding in interpreting graphs was found in The Most Production item. Figure 7.2 shows Farah's response based on her daily related interpretation of the word 'most'. She interpreted the 'most' production as the 'largest' number of shoes produced. During the interview, Farah pointed at the circle sign at the end of the solid line, as she thought it was the peak of production, meaning that the most shoes were produced at that time. She barely understood 'most' production to be the largest number of shoes produced over a period.

Figure 7.2

Farah's Response to Find the Time of The Most Production



Garis penuh (——) pada diagram menunjukkan jumlah sepatu yang diproduksi di sebuah industri rumah tangga pada suatu hari.

Pada selang waktu kapankah sepatu paling banyak diproduksi? Jelaskan dan hubungkan jawabanmu dengan diagram garis *produksi sepatu* di atas!

Pada jam 17.00, karena puncak pada diagram itu pada jam 17.00 menunjukkan angka 500

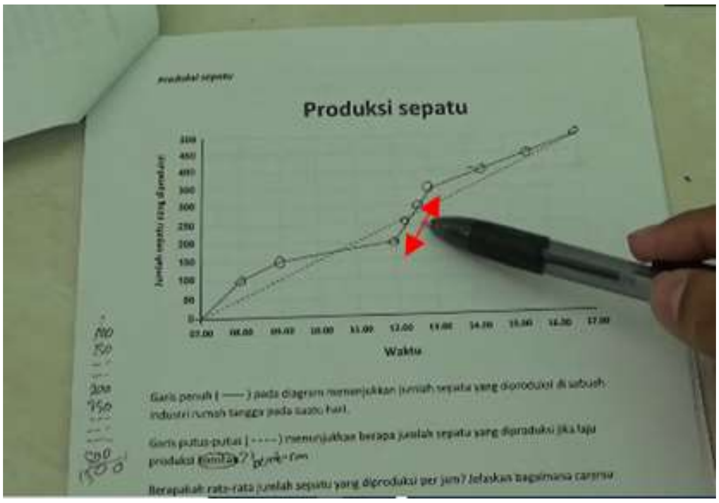
At 17.00, because the peak of the graph is at time 17.00 showing a value of 500

Similar to Farah, Cakra also used the idea of producing the largest number of shoes. However, Cakra added an imaginary context of production after lunchtime. During the interview, Cakra provided more details on his thoughts and added information not captured in his written response. For example, when asked to explain what he understood from the problem, he was sure the problem asked about the mode. This was not previously predicted, and some follow-up questions were posed for further investigation.

Three major ideas were summarised from Cakra's explanation during the interview: mode, rapid increase and the largest number. As with other students, Cakra was first asked to explain what he understood from the problem. Cakra stated that the mode is what this problem asked for, pointing to the solid line at 12.00 and 13.00 and stating that a rapid increase occurred there (see Figure 7.3). It was surprising that Cakra came up with the idea of this rapid increase during those hours as if he understood what the question asked and the line graph convention. However, after he was asked to explain further, he jumped to his previous thoughts on mode. He pointed to the end of the solid line and mentioned that it was the mode he was looking for (see Figure 7.4).

Figure 7.3

Cakra Showing the Rapid Increase



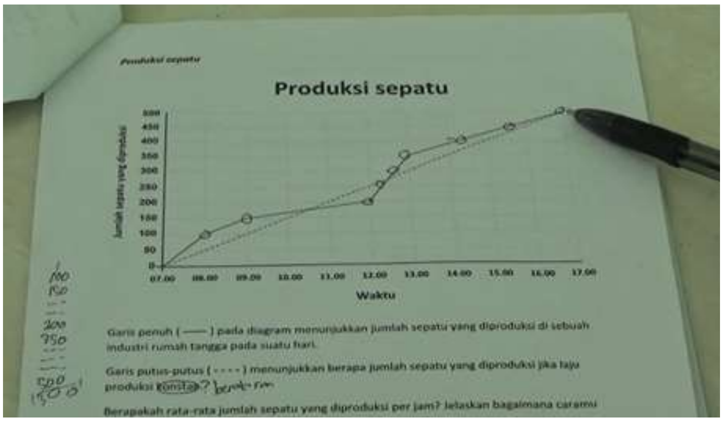
Garis penuh (——) pada diagram menunjukkan jumlah sepatu yang diproduksi di sebuah industri rumah tangga pada suatu hari.

Garis putus-putus (- - - -) menunjukkan berapa jumlah sepatu yang diproduksi jika laju produksi konstan? berubah-rubah

Berapakah rata-rata jumlah sepatu yang diproduksi per jam? Jelaskan bagaimana caramu

Figure 7.4

Cakra Showing the Highest Point as the Mode



Garis penuh (——) pada diagram menunjukkan jumlah sepatu yang diproduksi di sebuah industri rumah tangga pada suatu hari.

Garis putus-putus (- - - -) menunjukkan berapa jumlah sepatu yang diproduksi jika laju produksi konstan? berubah-rubah

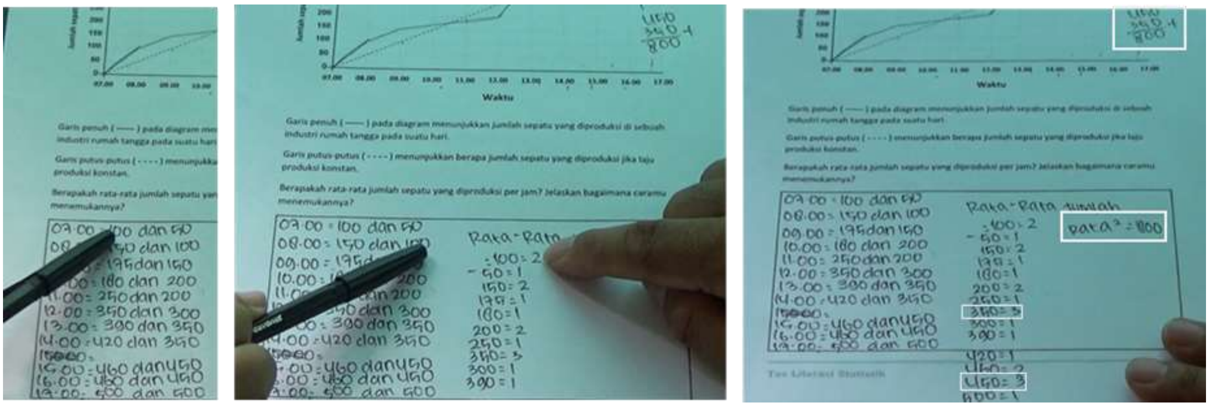
Berapakah rata-rata jumlah sepatu yang diproduksi per jam? Jelaskan bagaimana caramu

He was then asked to clarify his written answer with a probing question, ‘Which one was your answer, at 17.00 or at times after 12.00?’ In responding to this question, he explained that his answer was at 17.00, while at times after 12.00 was the explanation he provided on why the production rapidly increased because the employees had just finished lunch and their energy was restored. Considering his clarification, the idea of the largest number influenced him to find the answer for The Most Production item. His lack of understanding of the context of most production, the idea of an increase in the line graph, the line graph’s convention and the mode concept led him to an informal interpretation.

Finally, Ayu’s work exemplifies the Level 1 response as she used idiosyncratic thinking when interpreting both The Production Mean and The Most Production items. Ayu’s responses to both items reflected her absence of contextual, graphical and statistical understanding. To answer The Production Mean item (see Figure 7.5), she started to create a list of shoes produced each hour. However, she made a significant error from the very start by identifying that at 07.00, 100 (from the solid line) and 50 (from the dotted line) shoes were produced. This error results in incorrect data listings. Moreover, she counted the frequency from her list and chose the numbers with the highest frequency to be added up to result in the mean.

Figure 7.5

Ayu’s Work for The Production Mean Item



Garis penuh (——) pada diagram menunjukkan jumlah sepatu yang diproduksi di sebuah industri rumah tangga pada suatu hari.

Garis putus-putus (- - - -) menunjukkan berapa jumlah sepatu yang diproduksi jika laju produksi konstan.

Berapakah rata-rata jumlah sepatu yang diproduksi per jam? Jelaskan bagaimana caramu menemukannya?

07.00 = 100 dan 50
08.00 = 150 dan 100
09.00 = 195 dan 150
10.00 = 180 dan 200
11.00 = 250 dan 200
12.00 = 350 dan 300
13.00 = 390 dan 350
14.00 = 420 dan 350
~~15.00~~ =
15.00 = 460 dan 450
16.00 = 460 dan 490
17.00 = 500 dan 500

Rata-Rata Jumlah
100 = 2
50 = 1
150 = 2
195 = 1
180 = 1
200 = 2
250 = 1
350 = 3
300 = 1
390 = 1
420 = 1
460 = 2
450 = 3
500 = 1

rata² = 800

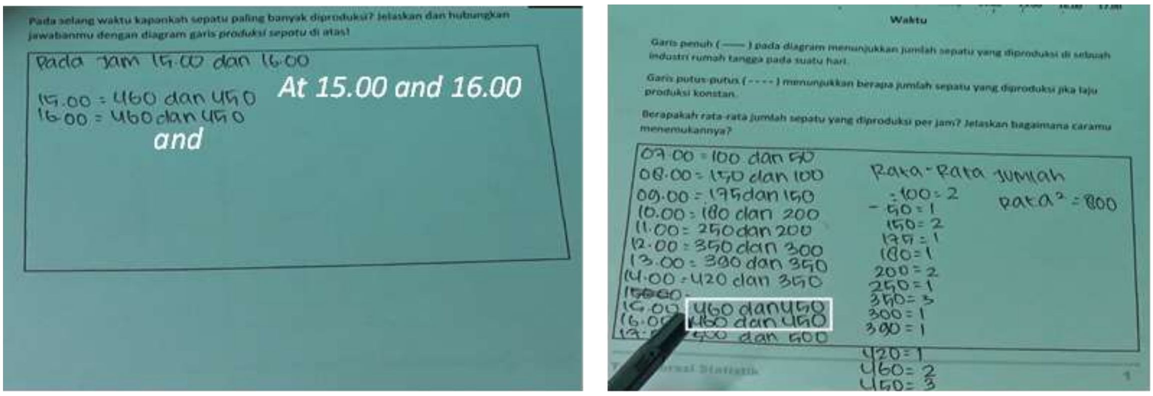
450
350 +
800

Tes Literasi Statistik

Her interview data revealed that Ayu (A) noticed the solid and dotted lines of the graph but could not explain the differences between the two lines. She knew the first item asked her about the mean, but she had no conceptual or procedural understanding of it. This idiosyncratic thinking by Ayu continued to appear in The Most Production item, as she interpreted the most production with the idea of 'the same production' (see Figure 7.6). Consequently, Ayu concentrated on locating 'the time intervals' with a 'similar' number of shoes. This was expressed during the interview when she pointed to times 15.00 and 16.00, as those times showed a similar number of shoes, 460 and 450 for both. This similarity was caused by her confusion when recording the number of shoes at 16.00, copying the number of shoes produced at 15.00.

Figure 7.6

Ayu's Work for The Most Production Item



Pada selang waktu kapankah sepatu paling banyak diproduksi? Jelaskan dan hubungkan jawabanmu dengan diagram garis produksi sepatu di atas!

Pada jam 15.00 dan 16.00

15.00 = 460 dan 450
16.00 = 460 dan 450

At 15.00 and 16.00

and

Waktu

Garis penuh (——) pada diagram menunjukkan jumlah sepatu yang diproduksi di sebuah industri rumah tangga pada suatu hari.

Garis putus-putus (- - - -) menunjukkan berapa jumlah sepatu yang diproduksi jika laju produksi konstan.

Berapakah rata-rata jumlah sepatu yang diproduksi per jam? Jelaskan bagaimana caramu menemukannya?

07.00 = 100 dan 50
08.00 = 150 dan 100
09.00 = 175 dan 150
10.00 = 180 dan 200
11.00 = 250 dan 200
12.00 = 350 dan 300
13.00 = 390 dan 350
14.00 = 420 dan 350
15.00 -
16.00 460 dan 450
16.00 460 dan 450
17. 500 dan 500

Rata-Rata Jumlah
100 = 2
50 = 1
150 = 2
175 = 1
180 = 1
200 = 2
250 = 1
350 = 3
300 = 1
390 = 1
420 = 1
460 = 2
450 = 3

Rata² = 800

I: *Can you please explain to me, what did you understand about this problem?*

A: *Which time interval.*

I: *What do you understand of 'at which time interval'?*

A: *It means similar.*

I: *Similar to what?*

A: *The number of shoes is the same.*

Look at my writing [for The Production Mean item]. *The similarity was at 15.00 and 16.00. Both have 460* [for solid line] *and 450* [for dotted line].

In addition to the abovementioned challenges in interpreting statistical information, the students also found challenges to effectively communicate data information to others based on their comprehension. From the interviewed students, it was determined that eight students performed at Level 3 on two communicating items and one student performed at Level 1. Among eight students performing at Level 3 on items assessing communicating skills, five of them were found in the YouTube Viewers item and three of them were found in the Domestic Waste item. These Level 3 students found challenges in selecting important components to be included in their summaries. These students showed insufficient knowledge of context, bar graph convention, statistics and mathematics ideas to include in their summaries. It was found that they all tend to provide data listings based on bar order as the important summary of information, regardless the different contexts. Moreover, some of them did not notice the graph convention such as the y-axis title that indicate the number was in thousands, as found in YouTube Viewers item. This response was found in Ayu (A), for example, who tried to summarise important information from YouTube Viewers item. She thought that the important information from the bar graph of YouTube Viewers was simply the total number of four-band viewers. To do this, she had to find the total number of viewers for each band. This was reflected by her responses at the beginning and end of the interview, as follows.

At the beginning of the interview:

I: *When you see the bar graph, what important information can you read?*

A: *From this.* [pointing to the text above the bar graph and reading it aloud]

At the end of the interview:

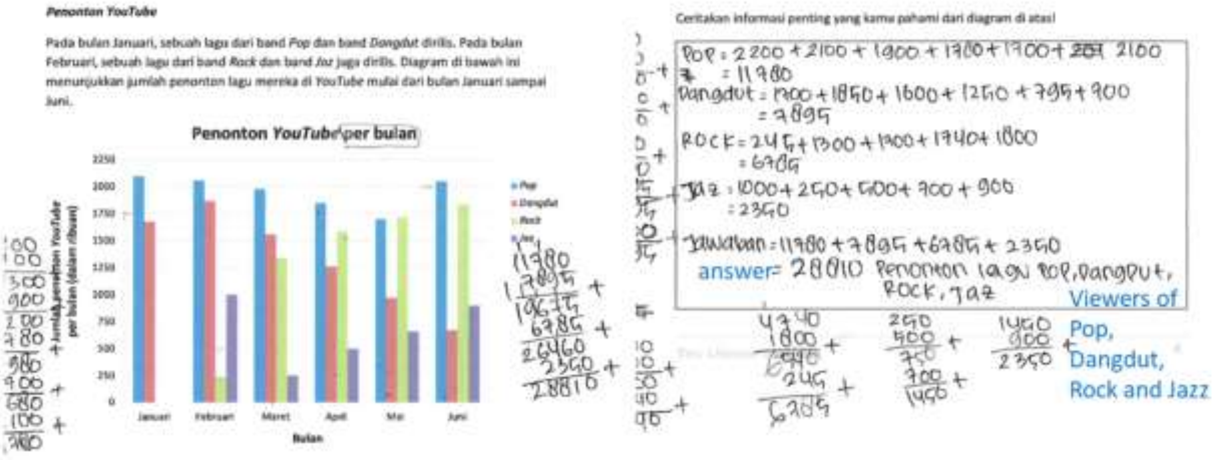
I: *After you got the number of viewers of Rock in each month, you added them all. Why did you add them?*

A: *Because this* [pointing to the question] *asked for [important] information.*

Ayu’s written response captures her answering process from start to end (see Figure 7.7). She began with finding the total number of viewers from January to June for each band, starting with *Pop*, *Dangdut*, *Rock* and *Jazz*. However, she did not notice that the number of viewers was presented in thousands. For example, she wrote ‘*Pop* = 2,200 + 2,100 + 1,900 + 1,780 + 1,700 + 2,100 = 11,780’ without adding further details that those numbers were in thousands. This implies that she did not pay attention to the y-axis title. Further, she tended to make inaccurate estimations as she focused on the number of viewers each month. For instance, Ayu wrote that *Pop* had 2,200 viewers in January, although its bar was closer to 2,000 than 2,250. This indicates that Ayu was unable to perform measurements using prediction or number line estimation. After determining the total number of viewers for each band, Ayu calculated the overall number of viewers for all bands, which she believed to be important information to share.

Figure 7.7

Ayu’s Written Response for YouTube Viewers Item



Penonton YouTube

Pada bulan Januari, sebuah lagu dari band *Pop* dan band *Dangdut* dirilis. Pada bulan Februari, sebuah lagu dari band *Rock* dan band *Jaz* juga dirilis. Diagram di bawah ini menunjukkan jumlah penonton lagu mereka di YouTube mulai dari bulan Januari sampai Juni.

Penonton YouTube per bulan
Jumlah penonton YouTube per bulan (dalam ribuan)
2250
2000
1750
1500
1250
1000
750
500
250
0
Januari
Februari
Maret
April
Mei
Juni
Bulan
Pop
Dangdut
Rock
Jaz
11780
17895
19675
6785
26460
2350
28810

Ceritakan informasi penting yang kamu pahami dari diagram di atas!

Pop = 2200 + 2100 + 1900 + 1780 + 1700 + ~~201~~ 2100
= 11780

Dangdut = 1700 + 1850 + 1600 + 1250 + 795 + 900
= 7895

ROCK = 245 + 1300 + 1900 + 1740 + 1800
= 6785

Jaz = 1000 + 250 + 500 + 700 + 900
= 2350

Jawaban = 11780 + 7895 + 6785 + 2350

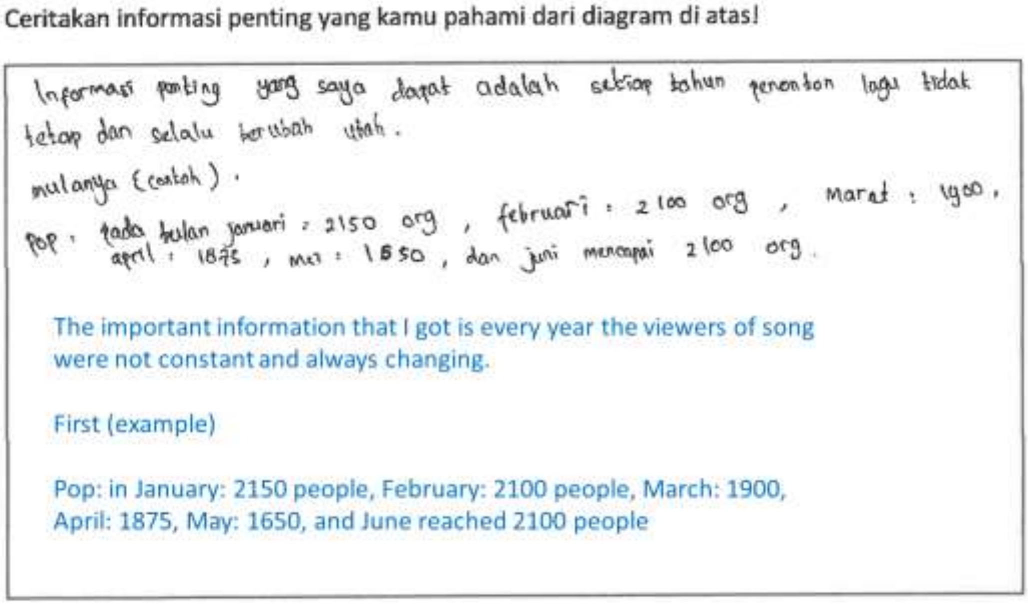
answer= 28810 Penonton lagu POP, Dangdut, ROCK, Jaz

Viewers of Pop, Dangdut, Rock and Jazz

Budi, another Level 3 student, attempted to compare the data to summarise the information presented in the bar graph. However, his comparison was inconsistent or, in other words, partly correct. The work of Budi in the YouTube Viewers item—presented in Figure 7.8—is exemplified to show the commonalities and differences with Ayu’s response.

Figure 7.8

Budi's Written Work for YouTube Viewers Item



Ceritakan informasi penting yang kamu pahami dari diagram di atas!

Informasi penting yang saya dapat adalah setiap tahun penonton lagu tidak tetap dan selalu berubah ubah.

mulanya (contoh).

Pop : pada bulan januari : 2150 org , februari : 2100 org , Maret : 1900 , april : 1875 , mei : 1650 , dan juni mencapai 2100 org .

The important information that I got is every year the viewers of song were not constant and always changing.

First (example)

Pop: in January: 2150 people, February: 2100 people, March: 1900, April: 1875, May: 1650, and June reached 2100 people

From Figure 7.8, it can be inferred that Budi started with a summary statement containing a comparison. His statement indicated that his comparison was based on data for each band, that he referred as a song. He merely compared the data for each band across the months rather than among the bands. In addition, his comparison only ended with changing numbers rather than trends. Moreover, he did not notice that the number of viewers presented in the table was in thousands. Thus, it can be summarised that Budi was able to make comparisons, but his limited understanding of graphs and statistical knowledge prevented him from including crucial information.

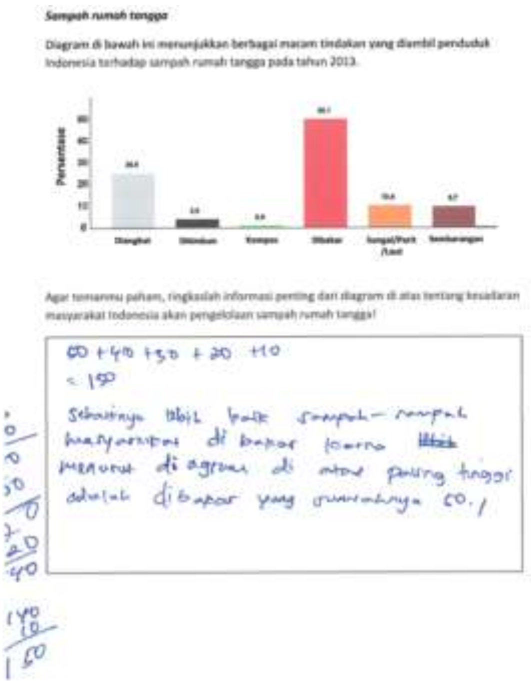
From the abovementioned description, showing the relationship between the data presented in bar graphs within the context is a challenge for Level 3 students. Ayu, Budi and the other Level 3 students did not include the most significant information they needed to report. Although they started comparing data, their summary information was very simple, such as ordering the data without showing its relationship. Based on this characteristic of

Level 3 students' challenges, students in Level 1 and 2 might only provide numerical information by merely listing the values in the bar graph.

In the case of communicating irrelevant information from the bar graph, the way Inggrid communicated information—categorised a Level 1—was exemplified (see Figure 7.9). Her written response to the Domestic Waste item reflected her absence of contextual, graphical and statistical-mathematical understanding. First, she did not completely understand the context, as she did not make any reference to people's awareness of domestic waste management. Second, she thought the number above the bar graph was the number of people instead of the percentage. This implies she did not pay attention to the y-axis title when interpreting the bars. Third, she added all the numbers from the y-axis, '50 + 40 + 30 + 20 + 10 = 150', without any further details about what it was used for. Finally, she suggested that all people should burn their domestic waste, as suggested by the data that burning represented the highest percentage in the bar graph. It seems she thought that the highest bar was always the answer.

Figure 7.9

Inggrid's Idiosyncratic Response When Asked to Communicate Important Information



Sampah rumah tangga

Diagram di bawah ini menunjukkan berbagai macam tindakan yang diambil penduduk Indonesia terhadap sampah rumah tangga pada tahun 2013.

Persentase
0
10
20
30
40
50
Diangkut
Ditimbun
Kompos
Dibakar
Sungai/Parit/Laut
Sembarangan

Agar temanmu paham, ringkaslah informasi penting dari diagram di atas tentang kesadaran masyarakat Indonesia akan pengelolaan sampah rumah tangga!

50 + 40 + 30 + 20 + 10
= 150

Sebaiknya lebih baik sampah-sampah masyarakat di bakar karena ~~lebih~~ menurut diagram di atas paling tinggi adalah dibakar yang jumlahnya 50.1

[illegible]

140
10
150

50+40+30+20+10
=150

Advisable, the waste from the people is burned, because ~~more~~ according to the graph above, the highest is burned which equals to 50.1

7.2.2 Students' Challenges in Evaluating Data-Based Claim and Making Decisions

Analysis of students' challenges in evaluating data-based claims revealed that their limited appreciation of the three components hindered them from providing relevant evidence as part of their argument. Similarly, analysis of students' challenges in decision-making revealed that their limited appreciation of the three components hindered them from providing relevant evidence to support their choice. It was found that the highest percentage of those students encountering challenges were in Level 3 for these skills (see again Table 7.2; for the larger cohort, see again Table 6.2). Further, based on data previously presented in Table 7.3, 13 responses in total were found at Level 3 across two evaluating items and nine responses were found in Level 1 and 2. In comparison, there were 13 responses in total performing at Level 3 across two decision-making items and eight responses were at Level 2. The Level 3 students showed inconsistent engagement with the context, tended to interpret the graphical details or tabular details rather than the context embedded and made conclusions without being accompanied by suitable statistical or mathematical justifications. Given these characteristics, it was determined that Level 3 students supported the existing claim, despite the item providing evidence to the contrary. In addition, they could not find relevant evidence to support their choice, in making decisions. Table 7.8 summarises Level 3 students' challenges in comprehending the three components of the evaluating items, resulting in their failure to contest the claim.

Table 7.8

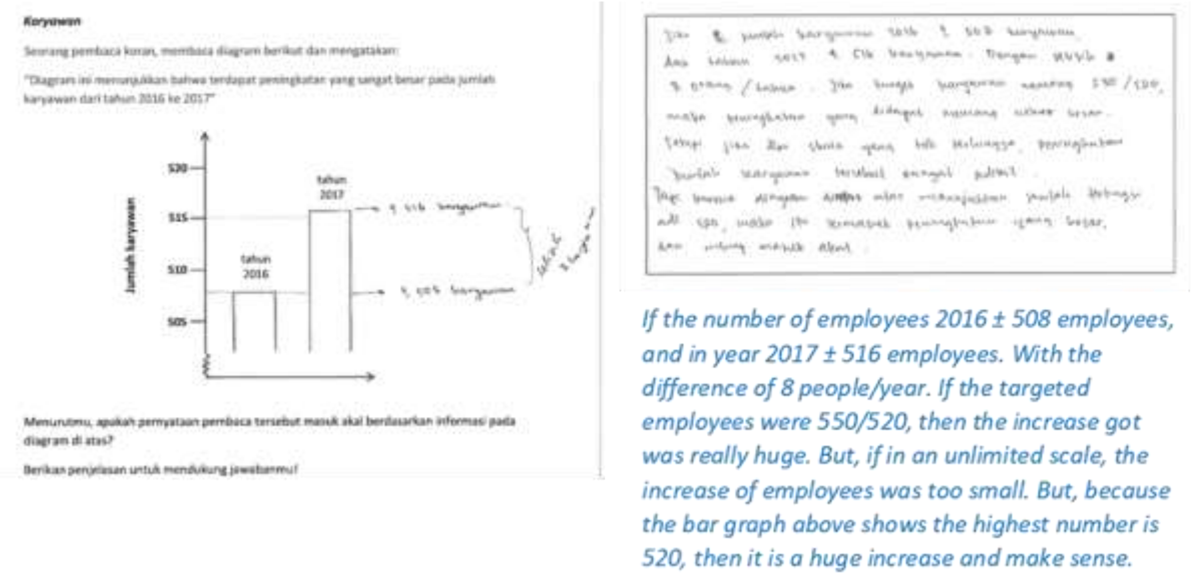
Level 3 Students' Challenges in Comprehending the Three Components for the Evaluating Items

	Text and context	Representation	Statistical-mathematical knowledge
The Employees	Student mixes up the meaning of 'huge increase' and 'increase'. Student merely uses key word 'huge increase' to make sense of the bar graph.	Student misunderstands the label on the axes and ignores the discontinuity in the y-axis. Student uses the difference of bars that almost double as justification.	Student could show or predict how many employees increase and use it to support the claim of a huge increase.
Mathematics Scores	Student ignores the provided mean and minimum passing grade.	Student knows that the y-axis shows the number of students but guess the meaning of the label in the x-axis.	Student recalculates mean from the bar graph. Student uses only intervals where class B is higher to support the claim.

With such inconsistent understanding, Level 3 students failed to find evidence from the bar graph to challenge the existing claim. For example, in the Mathematics Scores item, they could not utilise the information in the text about the minimum passing grade to further investigate and compare the number of students in both classes passing the test. Simply put, Level 3 students' understanding was insufficient to relate the context and data in the bar graph to reveal statistical and mathematical concepts in order to refute the claim. In The Employees item, they were distracted by the height of the two bars to show either an increase or a huge increase and ignored the discontinued y-axis. Figure 7.10 exemplifies the work of Inggrid, a Level 3 student, who supported the existing claim made by the newspaper reader for The Employees item as she failed to provide relevant evidence.

Figure 7.10

Inggrid's Irrelevant Evidence for The Employees Item



Karyawan

Seorang pembaca koran, membaca diagram berikut dan mengatakan:

"Diagram ini menunjukkan bahwa terdapat peningkatan yang sangat besar pada jumlah karyawan dari tahun 2016 ke 2017"

520
515
510
505
Jumlah karyawan
tahun 2016
tahun 2017

Menurutmu, apakah pernyataan pembaca tersebut masuk akal berdasarkan informasi pada diagram di atas?

Berikan penjelasan untuk mendukung jawabanmu!

If the number of employees 2016 ± 508 employees, and in year 2017 ± 516 employees. With the difference of 8 people/year. If the targeted employees were 550/520, then the increase got was really huge. But, if in an unlimited scale, the increase of employees was too small. But, because the bar graph above shows the highest number is 520, then it is a huge increase and make sense.

From Figure 7.10, it can be inferred that Inggrid started with evidence and concluded with her position, supporting the existing claim. In providing evidence, she calculated the difference in the number of employees (or the increase in the number of employees) between the two years (2016 and 2017), which she wrote as '8 people/year'. This is interesting, as she wrote 'per year' rather than only '8 people' as if it was a mean. In addition, she needed a benchmark to decide whether eight was a huge increase. Thus, she supposed that the benchmark was either 520 as shown on the y-axis, 550 or an undecided number above 550 to be the targeted employees. This implies that Inggrid had challenges in comprehending the meaning of the scale and label on the y-axis. As the bar graph showed 520, she concluded that the increase was huge and made sense. In other words, the increase to ±516 in 2017 was close to that of the targeted employees. The interview confirmed this interpretation of Inggrid's (Ig) thoughts.

I : *What about the question, what this problem asks for?*

Ig : *About… like… whether we agree or disagree with the [newspaper] reader.*

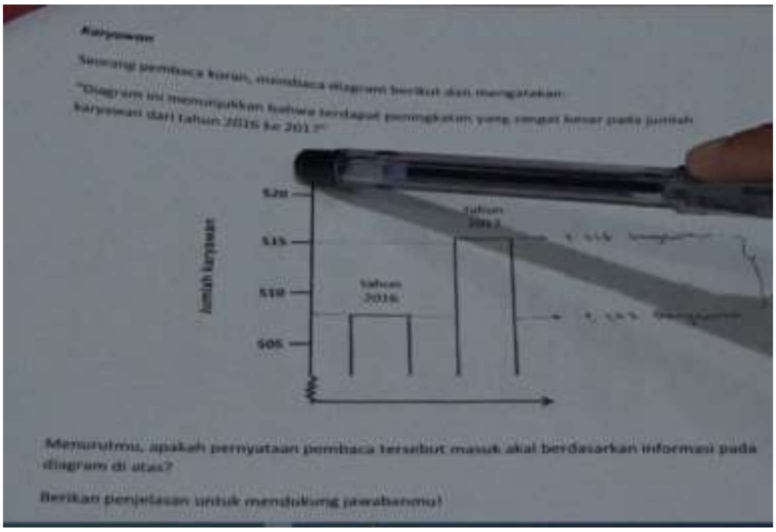
I : *And, when asked to agree or not, what did you think of after you looked at again the bar graph?*

Ig : *I myself somewhat disagree, but from the graph, 520 was the maximum. So, if we agree with the [newspaper] reader, that is right. I agree as the increase was quite huge. But if it was not from the graph, let's say if it was not limited to 520 or [if it was] far above, the increase was small.*

The interview confirmed that Inggrid had based her agreement and disagreement not on the proportion between the bars, which shows an almost twofold increase; instead, she based it on the maximum value in the y-axis, which is 520, as she interpreted it as the maximum targeted employees (see Figure 7.11). She noticed that the increase in 2017 was closest to 520.

Figure 7.11

Inggrid Pointing to 520 as Her Benchmark



Karyawan

Seorang pembaca koran, membaca diagram berikut dan mengatakan:

"Diagram ini menunjukkan bahwa terdapat peningkatan yang sangat besar pada jumlah karyawan dari tahun 2016 ke 2017"

Menurutmu, apakah pernyataan pembaca tersebut masuk akal berdasarkan informasi pada diagram di atas?

Berikan penjelasan untuk mendukung jawabanmu!

Inggrid's reasoning is an example of how Level 3 students used irrelevant evidence to support the existing claim because they found challenges in graph reading. Other students were found to be using similar irrelevant evidence. To further examine how students stated their position, Table 7.9 exemplifies some of the students' statements obtained from their written responses. Students who took the claim-supporting position seemed to have challenges finding the meaning of the context embedded in the graph. Some of them supported the claim as they ignored the word 'huge', and, consequently, considered a huge increase similar to an increase. However, others noticed the word 'huge' but agreed with the claim as they related to any particular benchmark. For example, Ucok did not differentiate between 'huge increase' and 'increase', while Putra noticed the word 'huge' but decided to agree with the claim as he reasoned it was about people, not other objects, and also because of the scale in the y-axis. Table 7.9 also illustrates students' irrelevant evidence to support their position; for example, Putra used the interval in the y-axis as a benchmark to decide that the increase is huge.

Table 7.9

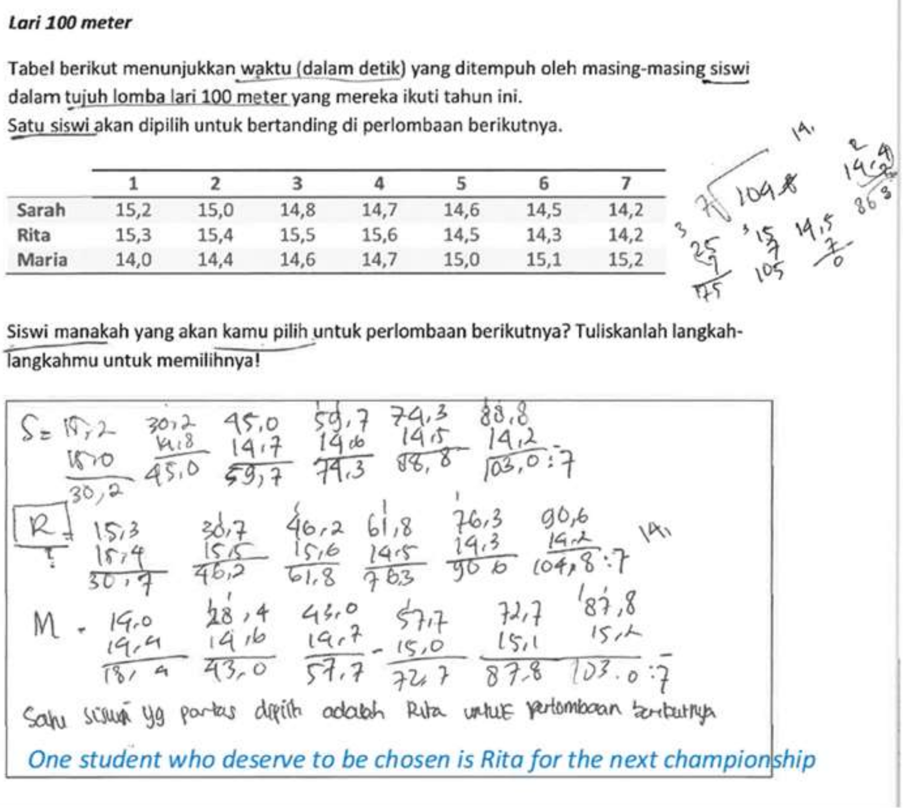
Students' Supporting Position and Irrelevant Evidence in Evaluating Claims

Position and evidence	Students' response
Students agree with the newspaper reader that there is an increase meaning a huge increase in the number of employees from 2016 to 2017	Inggrid: *[...],* then *it is a huge increase, and make sense.* Ucok: *[...] what the newspaper reader stated was correct because there was an increase in the number of employees.* Putra: *Yes, [...], a huge increase considering it is about people not objects.* Jesica: *[...]. So, the number of employees from 2016 to 2017 experienced an increase.*
Students irrelevant evidence	Inggrid: *Because the bar graph above shows the highest number is 520.* Putra: *The interval is 5 so that it has huge increase.*

Similar challenges in making sense of the three components were found when students solved decision-making items. Their challenges led them to make wrong decisions. For example, in The 100-Metre Race item, students' misunderstanding of the context of time in running competitions resulted in them choosing the best runner incorrectly. They may have devised the idea of comparing the mean of each runner's time, but they were found to choose the runner with the highest mean. Although the mean time spent by each runner was correctly calculated, choosing the runner with the highest mean was irrelevant. This evidence was typically provided by Level 3 students, who represented the highest proportion of the lower group. Figure 7.12 exemplifies the work of Ester, who used mean as her method to select the best runner for the upcoming championship.

Figure 7.12

Ester's Written Response for The 100-Metre Race Item



Lari 100 meter

Tabel berikut menunjukkan waktu (dalam detik) yang ditempuh oleh masing-masing siswi dalam tujuh lomba lari 100 meter yang mereka ikuti tahun ini.
Satu siswi akan dipilih untuk bertanding di perlombaan berikutnya.

	1	2	3	4	5	6	7
Sarah	15,2	15,0	14,8	14,7	14,6	14,5	14,2
Rita	15,3	15,4	15,5	15,6	14,5	14,3	14,2
Maria	14,0	14,4	14,6	14,7	15,0	15,1	15,2

14,
7√104,8
14,9
86 3
25 7 175
15 7 105
14,5 7 0

Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!

S = 15,2 + 15,0 = 30,2   30,2 + 14,8 = 45,0   45,0 + 14,7 = 59,7   59,7 + 14,6 = 74,3   74,3 + 14,5 = 88,8   88,8 + 14,2 = 103,0 : 7

R = 15,3 + 15,4 = 30,7   30,7 + 15,5 = 46,2   46,2 + 15,6 = 61,8   61,8 + 14,5 = 76,3   76,3 + 14,3 = 90,6   90,6 + 14,2 = 104,8 : 7   14,

M = 14,0 + 14,4 = 18,4   28,4 + 14,6 = 43,0   43,0 + 14,7 = 57,7   57,7 + 15,0 = 72,7   72,7 + 15,1 = 87,8   87,8 + 15,2 = 103,0 : 7

Satu siswa yg pantas dipilih adalah Rita untuk perlombaan berikutnya

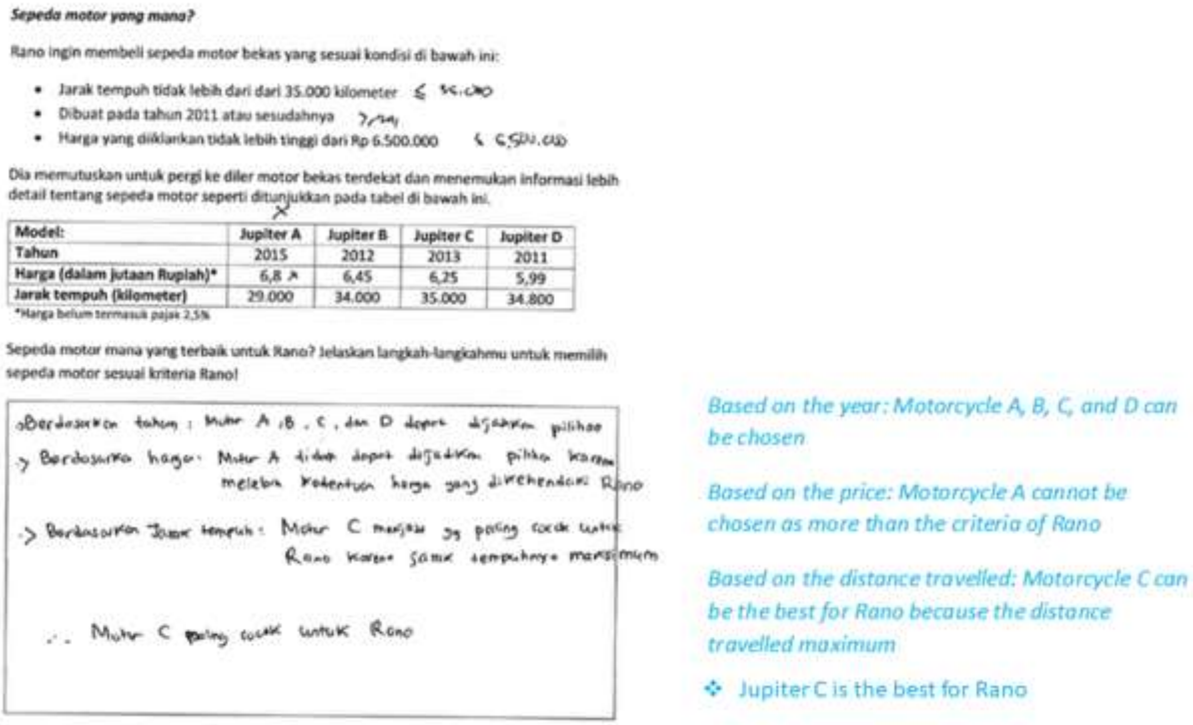
One student who deserve to be chosen is Rita for the next championship

Figure 7.12 shows that Ester used the mean of each runner's time across seven races as the selection method. She started by adding the time from races one to seven and dividing it by seven. The total time she found was correct for each runner. In addition, the mean time spent by each runner across the seven races was correctly calculated. However, she stated that Rita should be selected for the upcoming championship. Based on Ester's calculations, Rita's mean was the highest among the three runners; in actuality, this meant it was the slowest. Thus, it was interpreted that Ester had an incorrect interpretation of time in a running competition—she thought that the highest mean in a running competition determined the most likely winner.

Similarly, in the Which Motorcycle? item, Level 3 students did not understand the meaning of distance travelled or calculate the tax. Figure 7.13 illustrates Noval's irrelevant evidence for choosing a motorcycle that meets the three numerical conditions. His irrelevant evidence was influenced by his limited knowledge of the context, particularly regarding the distance travelled. It was clear that he understood the year criteria and that all four motorcycles met the year criteria. In terms of the distance travelled, however, Noval showed an understanding between informal and inconsistent levels. In particular, Noval incorrectly interpreted that the greatest distance travelled was the maximum distance a motorcycle could travel before running out of petrol, and thus he chose *Jupiter C* for Rano. In terms of price, Noval showed an understanding, but he did not include the tax. His exclusion of tax resulted in *Jupiter B* as one of the options. The interview with Noval (N) clarified what he understood, supporting the interpretation of his written responses. Noval's contextual misunderstanding of the distance travelled was explained during the interview.

Figure 7.13

Noval’s Irrelevant Evidence for His Choice for Which Motorcycle? Item



Sepeda motor yang mana?

Rano ingin membeli sepeda motor bekas yang sesuai kondisi di bawah ini:

- Jarak tempuh tidak lebih dari dari 35.000 kilometer ≤ 35.000
- Dibuat pada tahun 2011 atau sesudahnya ≥ 2011
- Harga yang diiklankan tidak lebih tinggi dari Rp 6.500.000 ≤ 6.500.000

Dia memutuskan untuk pergi ke diler motor bekas terdekat dan menemukan informasi lebih detail tentang sepeda motor seperti ditunjukkan pada tabel di bawah ini.

Model:	Jupiter A	Jupiter B	Jupiter C	Jupiter D
Tahun	2015	2012	2013	2011
Harga (dalam jutaan Rupiah)*	6,8 ↗	6,45	6,25	5,99
Jarak tempuh (kilometer)	29.000	34.000	35.000	34.800

*Harga belum termasuk pajak 2,5%

Sepeda motor mana yang terbaik untuk Rano? Jelaskan langkah-langkahmu untuk memilih sepeda motor sesuai kriteria Rano!

∘ Berdasarkan tahun : Motor A, B, C, dan D dapat dijadikan pilihan

-> Berdasarkan harga: Motor A tidak dapat dijadikan pilihan karena melebihi ketentuan harga yang dikehendaki Rano

-> Berdasarkan Jarak tempuh: Motor C menjadi yg paling cocok untuk Rano karena jarak tempuhnya maksimum

∴ Motor C paling cocok untuk Rano

Based on the year: Motorcycle A, B, C, and D can be chosen

Based on the price: Motorcycle A cannot be chosen as more than the criteria of Rano

Based on the distance travelled: Motorcycle C can be the best for Rano because the distance travelled maximum

- Jupiter C is the best for Rano

I: *What do those different distances travelled mean?*

N: *They show how far each motorcycle could travel before its petrol running out. The effectiveness of its machine.*

Further, the interview clarified his mathematical understanding of tax and price. Noval had a good sense of tax when it was added to the price, but he seemed to forget that the price, including the tax, should be tied to the criteria.

N: *I missed that tax. But this information was not too important because each motorcycle has this tax. Although the tax was counted, it would be the same. Jupiter A would be still the most expensive, and the rests would remain in the same order.*

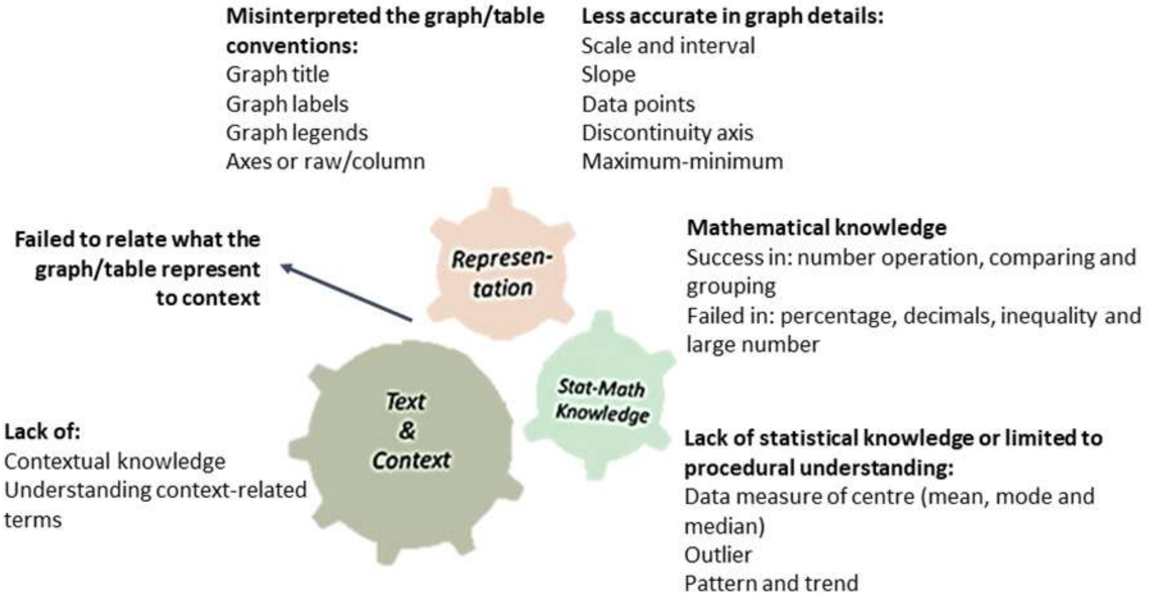
7.2.3 Summary of Students' Challenges

Students' challenges in interpreting, communicating, evaluating and making decisions were influenced by their challenges with the three components of SL. Because the three components are interrelated, students' challenges in one of the three components affect their challenges in the other two. The lower group of students (Levels 1–3) found challenges in making sense of the context embedded in the representation, the data in the representation and the statistical-mathematical knowledge that arises from the representation.

In particular, Level 3 students struggle more with contextual-graphical interrelationship than with statistical-mathematical knowledge. Level 3 students demonstrate an initial understanding of statistical idea such as mean and its arithmetic procedure, but their comprehension is dependent on the types of representation and their familiarity with contexts. When interpreting the line graph, the lower group of students merely read the value, and if they performed the calculation, the result was procedurally correct but contextually incorrect. In communicating the most relevant data from the bar graph, these students could not choose the most significant features from the representation to report, and an inappropriate contextual explanation or numerical information was used. In evaluating the claim, students hardly found relevant statistical evidence from the graph to challenge the claim, resulting in them supporting the claim. Finally, in making decisions, students were unable to show relevant evidence to support their choice because they misinterpreted the context and statistical concepts such as mean, while having ability to execute calculations. Figure 7.14 summarises students' challenges in making sense of the three components of SL, causing them to fail to respond to statistical information based on the assessed skill.

Figure 7.14

Students' Challenges on the Three Components of SL


Misinterpreted the graph/table conventions:
Graph title
Graph labels
Graph legends
Axes or raw/column
Less accurate in graph details:
Scale and interval
Slope
Data points
Discontinuity axis
Maximum-minimum
Failed to relate what the graph/table represent to context
Represen-tation
Mathematical knowledge
Success in: number operation, comparing and grouping
Failed in: percentage, decimals, inequality and large number
Text & Context
Stat-Math Knowledge
Lack of:
Contextual knowledge
Understanding context-related terms
Lack of statistical knowledge or limited to procedural understanding:
Data measure of centre (mean, mode and median)
Outlier
Pattern and trend

7.3 Students' Understandings of the Three SL Components

In contrast to students' challenges, students in the upper group (Levels 4–6) showed appropriate or critical thinking in the three components. These upper group students' responses revealed information about their understandings of the three components of SL. To further reveal their understanding, the responses of the upper group students were analysed. Using their written responses and interviews data, more detailed understandings are presented in relation to the four skills assessed.

Subsequent sections will delve into these understandings in greater detail. These sections particularly focus on Level 5 and 6 responses to show students' critical thinking on the three components. This is essential given that the majority of students lack critical thinking. However, it is important to begin with Level 4 (consistent non-critical) responses to demonstrate students' appropriate understandings of the three components of SL and how they differ from the Levels 5-6 responses. Moreover, the highest percentage of students in the upper group was in this Level 4, suggesting this number of students may have challenges to advance to critical thinking.

7.3.1 Students' Appropriate to Critical Responses in Interpreting and Communicating Data-Based Information

The responses from students in Level 4 showed their appropriate understanding while the responses from students in Levels 5 and 6 showed their critical understanding when interpreting and communicating statistical information. Using the same interpreting items (The Production Mean and The Most Production) and communicating items (YouTube Viewers and Domestic Waste) to reveal students' challenges, their appropriate and critical understanding on the three SL components (text and context, representation and statistical-mathematical knowledge) were examined. Based on data previously presented in Table 7.3, nine responses were at Level 4 and 24 responses could be categorised to achieve the highest two levels for interpreting items. For communicating, 27 responses were found to achieve Level 4, 12 responses were found to achieve Level 5 and no response can achieve Level 6.

Among the students demonstrating appropriate responses, there were three students demonstrating consistency in providing appropriate responses across interpreting and communicating items. Table 7.10 presents these three students and their achieved levels on the four items. Examining their written and interview responses revealed insights into their appropriate understandings.

Table 7.10

Three Students Demonstrating Appropriate Responses in Interpreting and Communicating Statistical Information

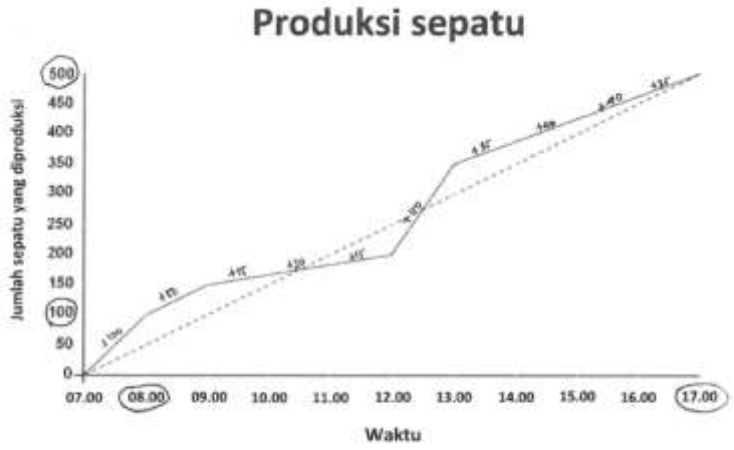
Students	Level			
	The Production Mean (I)	The Most Production (I)	YouTube Viewers (C)	Domestic Waste (C)
Hannah	Consistent non-critical	Consistent non-critical	Consistent non-critical	Consistent non-critical
Galang	Inconsistent	Consistent non-critical	Consistent non-critical	Consistent non-critical
Noval	Inconsistent	Consistent non-critical	Consistent non-critical	Consistent non-critical

Note. I is Interpreting and C is Communicating.

For interpreting, Level 4 students' appropriate understanding of the three components assisted them in solving, for example, The Production Mean item. Students at this Level 4 showed a non-critical understanding of the context of an increase in shoe production. Figure 7.15 shows Hannah's written work on the interpreting item, and it can be interpreted that she first identified what she needed to calculate the hourly mean of shoe production. She needed to find the total number of shoes produced and the total production time from the line graph. However, Hannah looked at the data in the line graph, particularly from the solid line, as if it was collected on an hourly basis. She could not understand that the data was collected at particular times instead of on an hourly basis. In addition, she initially struggled to understand the context of an increase but gradually could reflect it. In the process of determining the number of shoes produced per hour, she changed her thinking three times. Finally, it seemed she could show self-correction when looking at the data increase on an hourly basis.

Figure 7.15

Hannah's Appropriate Response to The Production Mean Item



Garis penuh (——) pada diagram menunjukkan jumlah sepatu yang diproduksi di sebuah industri rumah tangga pada suatu hari.

Garis putus-putus (- - - -) menunjukkan berapa jumlah sepatu yang diproduksi jika laju produksi konstan.

Berapakah rata-rata jumlah sepatu yang diproduksi per jam? Jelaskan bagaimana caramu menemukannya?

~~100, 150, 175,~~

~~100, 150, 165, 185, 200, 350, 385, 425~~

100, 50, 15, 20, 15, 150, 35, 40, 40, 35

mean = $\frac{500}{10}$ = 50

That interpretation of Hannah's written work was clarified by her interviews. During the interview, Hannah (H) showed an understanding of the question asked by saying '[its] mean per hour'. The 'per hour' basis seemed to have led her to find the number of shoes on an hourly basis. For that, she needed to make predictions for certain hours, such as at 10.00 and 11.00. Interestingly, her prediction of the number of shoes produced at 10.00 and 11.00 did not change the total number of shoes within two data points, from 9.00 to 12.00. This means she implicitly could recognise that the data points matter. As a result, she found the correct number of shoes across the 10 hours. Eventually, she applied the add-divide formula, dividing 500 by 10, resulting in 50, the mean number of shoes produced per hour.

I: *After comprehending the question, what did you think of?*

H: *Looking at each hour's number of shoes.*

I: *Can you please demonstrate how you interpret the line graph?*

H: *At 08.00, 100 [pairs of shoes] were produced; at 09.00, 150 [pairs of shoes] were produced; at 10.00, there were approximately 175, oops 165 [pairs of shoes] produced when looked from this* [pointing at the solid line at 10.00]*; at 11.00, there were approximately 185 [pairs of shoes] produced.*

She was then asked to explain why she changed her list.

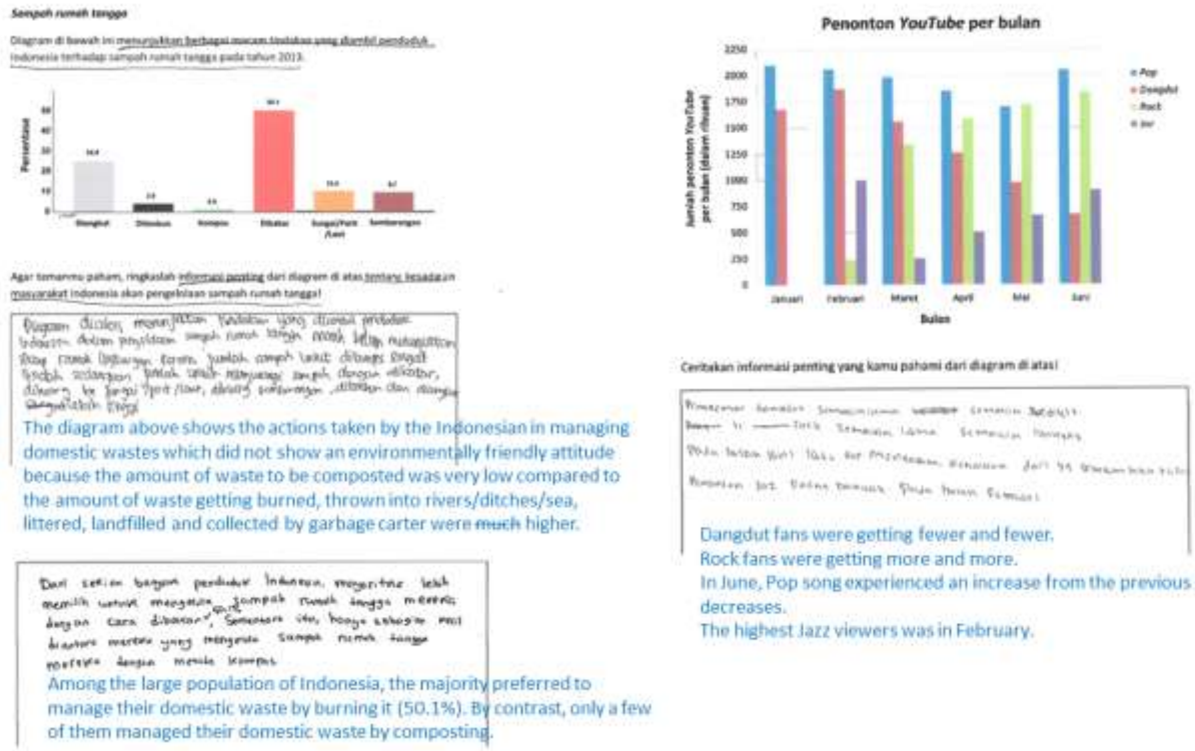
I: C*ould you tell me, what your initial thoughts for ~~100, 150, 175?~~*

H: *Initially, I believed it was like a number line, so I would begin at 100 and 150* [pointing at the solid line]. *After considering it, it seemed inappropriate. Then, I realised this one hour increases by 100* [pointing at the solid line between 07.00 and 08.00] *and the next one hour increases by 50* [pointing at the solid line between 08.00 and 09.00] *until this* [pointing at the end of the solid line]. *Then, I added them all and the total must be 500* [pointing at the 500 in y-axis] *which I divided by 10 [hours] showing the total time from 07.00 to 17.00. Therefore, 500 divided by 10 equals 50.*

Similarly, in communicating their views to data-based information, Level 4 students exhibited non-critical thinking. Figure 7.16 exemplifies this thinking of the three students: Hannah and Noval for Domestic Waste item and Galang for YouTube Viewers item. These three students demonstrated an understanding of the three SL components, but they lacked critical graphical competence and linguistic competence to communicate relevant information in effective ways. Although they already highlighted trends, made simple comparisons and identified certain significant data from data presented in the graphs, they did not include numeric information to support their summary. Moreover, these students missed in showing any possible relationships among variables in critical ways. The following is an explanation of their responses.

Figure 7.16

Examples of Appropriate Response in Communicating Data-Based Information


Persentase
50.1
Diangkut
Ditimbun
Kompos
Dibakar
Sungai/Parit/Laut
Sembarangan
Penonton YouTube per bulan
Jumlah penonton YouTube per bulan (dalam ribuan)
2250
2000
1750
1500
1250
1000
750
500
250
0
Januari
Februari
Maret
April
Mei
Juni
Bulan
Pop
Dangdut
Rock
Jazz

Sampah rumah tangga

Diagram di bawah ini menunjukkan berbagai macam tindakan yang diambil penduduk Indonesia terhadap sampah rumah tangga pada tahun 2013.

Agar temanmu paham, ringkaslah informasi penting dari diagram di atas tentang kesadaran masyarakat Indonesia akan pengelolaan sampah rumah tangga!

Diagram diatas, menunjukkan tindakan yang diambil penduduk Indonesia dalam pengelolaan sampah rumah tangga [illegible] [illegible] [illegible] [illegible] karena jumlah sampah untuk dibuat kompos sedikit sedangkan jumlah untuk mengurangi sampah dengan dibakar, dibuang ke sungai/parit/laut, dibuang sembarangan, ditimbun dan diangkut [illegible] tinggi

The diagram above shows the actions taken by the Indonesian in managing domestic wastes which did not show an environmentally friendly attitude because the amount of waste to be composted was very low compared to the amount of waste getting burned, thrown into rivers/ditches/sea, littered, landfilled and collected by garbage carter were ~~much~~ higher.

Dari sekian banyak penduduk Indonesia, mayoritas lebih memilih untuk mengatasi sampah rumah tangga mereka dengan cara dibakar, sementara itu, hanya sebagian kecil diantara mereka yang mengatasi sampah rumah tangga mereka dengan membuat kompos

Among the large population of Indonesia, the majority preferred to manage their domestic waste by burning it (50.1%). By contrast, only a few of them managed their domestic waste by composting.

Ceritakan informasi penting yang kamu pahami dari diagram di atas!

Penggemar dangdut semakin ~~[illegible]~~ semakin sedikit
~~[illegible]~~ [illegible] Rock semakin lama semakin banyak
Pada bulan Juni lagu Pop mengalami kenaikan dari yang sebelumnya turun
Penonton Jazz paling banyak pada bulan Februari

Dangdut fans were getting fewer and fewer.
Rock fans were getting more and more.
In June, Pop song experienced an increase from the previous decreases.
The highest Jazz viewers was in February.

Note. The figure at the top left depicts Hannah's written response; the figure in the bottom left depicts Noval's written response; the figure on the right depicts Galang's written response.

The Level 4 students demonstrated proficiency in identifying certain significant data from graphs as important information to share. For example, Hannah could identify the minimum data (Domestic Waste), Noval could identify the maximum and minimum data to be contrasted (Domestic Waste) and Galang could recognise the decreasing trend in Dangdut viewers and the increasing trend in Rock viewers (YouTube Viewers). Among these three students, Hannah appeared to be thinking of a slightly different level of understanding as she addressed the question on how aware the public is concerning waste management. She said in the beginning of her response that Indonesian people “did not show environmentally friendly attitude because the amount of wastes to be composted was very low”. This kind of statement was a good summary but did not appear in the other two students or most of Level 4 students.

The typical responses from Level 4 students were primarily descriptive based on the graph visual. Although their responses were appropriate, such responses lacked critical language competence and numerical evidence. In the case of Hannah’s response for Domestic Waste item, her response only compared composting with the other waste managements. It might imply that composting is the only acceptable method of managing waste, overlooking landfill and garbage carter. In the case of Noval’s and Galang’s responses, they both appeared to ‘compare’ but in different ways. Noval wrote sentences that highlighted the comparison of minimum and maximum data without emphasising the reasons behind, while Galang succinctly listed the trend for each band without providing any further context. Above all, the three students did not include any numerical data to support their comparison. The only numerical evidence was found in Noval’s response, but it was insufficient. Their interview data emphasised the above interpretation of their written works.

An interview with Hannah (H)

H: *I think the best way to reduce waste is by composting.*

I: *Is composting the only waste management that is good for environment?*

H: *Yes.*

An interview with Noval (N)

I: *What does this question ask for?*

N: *This question asks for a summary of the important information from the graph about the Indonesian people’s awareness of domestic waste management.*

I: *What important information that you could capture when you saw this bar graph?*

N: *I only looked at the bar with the highest percentage and bar with the smallest percentage.*

An interview with Galang (G)

I: *What does this question ask for?*

N: *It is about the important information from the graph*

I: *When you looked at the graph, what important information you can tell.*

N: *The [number of] Dangdut's fans was getting fewer over the time.*

By comparison, the critical understanding demonstrated by Level 5 and 6 students differ from those of Level 4 students. Table 7.11 presents the summary of students' critical thinking in interpreting and communicating information in relation to the three SL components.

Table 7.11

Students' Critical Interpretation and Communication Skills in Relation to Three SL Components

Components	Students' critical understanding
Text and context	Students demonstrate critical understandings of the textual and contextual information. These textual understandings can be applied to various contexts, including entertainment, economic and environmental. These contexts provide students with strategies and procedures for interpreting and communicating data-based information. Students understand, for instance, that the text used to explain two lines in The Production Mean and The Most Production items represent the same data: the solid line represents the raw data, and the dotted line represents the processed data. Students also know when the data was collected and the difference of proper versus improper waste managements (Domestic Waste item) and when the bands released singles (YouTube Viewers item).

Table 7.11 (continued)

Components	Students' critical understanding
Representation	The capacity of students to interpret and communicate data on the representation is not restricted by the types of graphs. They demonstrate a sophisticated comprehension of both line and bar graph conventions. First, students thoroughly investigate the graph's features (i.e., title, axes, headings and legends). They understand, for example, that the number of YouTube viewers displayed on the y-axis was in thousands (e.g., 250 means 250 thousand viewers) and that the legend depicts the four bands in various colours. Second, students effectively determine what the numbers represent (e.g., by searching for the largest and smallest values in one or more categories to get a sense of the data). Students, for example, contrast the highest (burning) and lowest (composting) bar in the Domestic Waste item as improper and appropriate waste management. Thirdly, students identify the differences in values (e.g., the changes of data over time and the comparative values of data within a category). For instance, students correctly identify the increase of shoe production over the time. Fourth, students successfully identify where differences exist (e.g., using information from Step 3 to make comparisons between two or more categories or time periods). Using the steepest slope, for instance, students were able to identify the time interval with the highest production. Finally, students were able to assess why those differences occurred by searching for reasons for the relationships in the data and making connections to the context. For instance, students were able to identify a trend over time and relate it to the context.
Statistical-mathematical knowledge	Students' knowledge of statistical concept and related mathematics procedures and concept are critical and thus enabled them to correctly interpret and communicate the numbers used in statistical reports. They, as data consumers, know the functions of statistical concepts more than the underlying calculations. They show flexibility in using average and trend in their responses. For example, students identify the overall trend for each band to be compared in the YouTube Viewers and make a group of the bands based on the same pattern or trend. Additionally, students demonstrate sufficient level of mathematical knowledge on rational numbers such as percentages, fractions, totals, proportions and on number operations, especially addition and division. They use their mathematical knowledge to compare the most and least watched bands (YouTube Viewers), determine the most production using proportional comparison (The Production Mean), compare the most and least methods of waste management (Domestic Waste), make a group of the actions into proper and improper actions to compare the percentages (Domestic Waste) and change the percentage to obtain additional information (e.g., 50.1% means more than a half and 0.8 means less than 1%).

Among the students demonstrating critical responses, there were four students demonstrating consistency in providing critical responses across interpreting and communicating items. Table 7.12 presents these four students and their achieved levels on four items. Comparing their written responses revealed insights into their critical understandings.

Table 7.12

Four Students Demonstrating Critical Responses in Interpreting and Communicating Statistical Information

Students	Level			
	The Production Mean (I)	The Most Production (I)	YouTube Viewers (C)	Domestic Waste (C)
Jesica	Critical mathematical	Critical	Consistent non-critical	Critical
Luhut	Critical mathematical	Critical mathematical	Critical	Critical
Ucok	Critical mathematical	Critical mathematical	Consistent non-critical	Critical
Xavier	Critical mathematical	Critical mathematical	Critical	Consistent non-critical

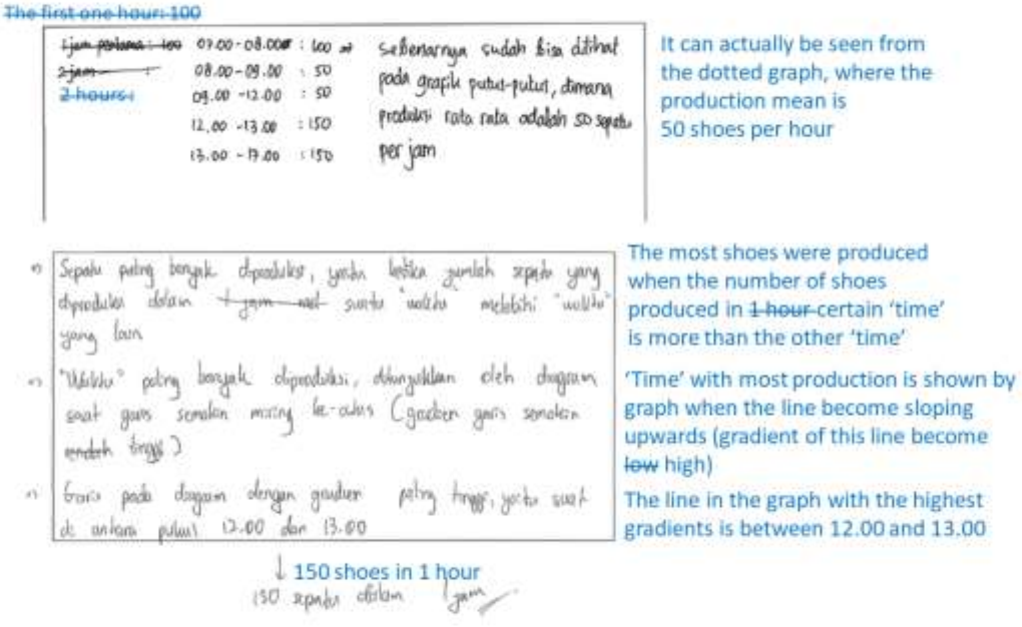
Note. I is Interpreting and C is Communicating.

For interpreting, Level 5 students' critical understanding of the three components assisted them in efficiently solving, for example, The Production Mean item. Students at this Level 5 showed a critical understanding of the context of an increase in shoe production. With such contextual understanding, Level 5 students created a list of the number of shoes produced based on the data points rather than each hour. This indicated that Level 5 students were able to recognise that the data displayed in the solid line was not collected hourly. In other words, Level 5 students demonstrated critical contextual and graphical understandings. Even though they ignored the dotted line and solely focused on the solid line, this would not be a major problem because they understood the concept of an increase. After determining the total number of shoes produced and the total production time, Level 5 students employed the mean formula using the idea of add-divide procedure to determine the hourly mean of shoe production.

Comparatively, Level 6 students successfully comprehended the relationship between the context and the represented data. They required no assistance from the increase shown by the solid line to determine the total number of shoes produced during the provided times. Instead, they could ascertain the total number of shoes over time using the relationship between the y-axis and x-axis. In addition, they demonstrated the ability to focus on either the solid line or the dotted line which shows a constant increase over the period. These understandings enabled them to determine the *mean* with procedure and with graph display (for The Production Mean item) or the *mode* with proportional comparison and graph characteristics (for The Most Production item). Figure 7.17 exemplifies how the interconnected comprehension of these Level 6 students regarding the three components was reflected in their written responses.

Figure 7.17

Xavier's and Luhut's Critical Interpreting Skill



~~The first one hour: 100~~

~~1 jam pertama : 100~~ 07.00 - 08.00 : 100
~~2 jam~~ 08.00 - 09.00 : 50
~~2 hours :~~ 09.00 - 12.00 : 50
12.00 - 13.00 : 150
13.00 - 17.00 : 150

Sebenarnya sudah bisa dilihat pada grafik putus-putus, dimana produksi rata rata adalah 50 sepatu per jam

It can actually be seen from the dotted graph, where the production mean is 50 shoes per hour

Sepatu paling banyak diproduksi, yaitu ketika jumlah sepatu yang diproduksi dalam ~~1 jam~~ suatu "waktu" melebihi "waktu" yang lain

The most shoes were produced when the number of shoes produced in ~~1 hour~~ certain 'time' is more than the other 'time'

"Waktu" paling banyak diproduksi, ditunjukkan oleh diagram saat garis semakin miring ke-atas (gradien garis semakin ~~rendah~~ tinggi)

'Time' with most production is shown by graph when the line become sloping upwards (gradient of this line become ~~low~~ high)

Garis pada diagram dengan gradien paling tinggi, yaitu saat di antara pukul 12.00 dan 13.00

The line in the graph with the highest gradients is between 12.00 and 13.00

150 shoes in 1 hour
150 sepatu dalam 1 jam

Note. The above figure is Xavier's written response for The Production Mean item; The below figure is Luhut's written response for The Most Production item.

Both Xavier and Luhut exhibited critical understanding of the three SL components, as depicted in Figure 7.17. It was conjectured that Xavier began with the listing based on the solid line and ended with the statements based on the dotted line. Initially, he identified what was necessary for calculating the hourly mean of shoe production. From the line graph, he needed to determine the total number of shoes produced and the total production time. Realizing that the data had been collected at specific point of times rather than hourly, he compiled a list of the number of shoes in each data point. However, he did not continue counting the mean using formula; rather, he concluded by stating 'As can be seen from the graph's dotted line, the mean production is 50 per hour'. It indicated that he correctly interpreted the constant hourly increase in the dotted line as the mean. Similarly, Luhut successfully interpreted the steepest slope as the highest production, focusing on the time interval between 12.00-13.00 with the steepest slope.

Through their interviews, the interpretation of Xavier's written work was clarified. Xavier (X) explained during the interview that his list was obtained from the data points on the solid line. He confidently explained his list by referring to the solid line, which indicated that 100 shoes were produced in the first one hour, 50 shoes were produced in the second hour, and 150 shoes were produced in the following three hours. He subsequently carried out mental calculation to determine the mean production, by determining first the total number of shoes and total hours. Instead of writing it, he alternatively looked at the dotted line. He confirmed himself that the constant increase of the dotted line indicates the mean production. His interview demonstrated that Xavier has critical understanding of the three SL components: textual and contextual understanding, as evidenced by his comprehension of the production increase; graphical understanding, as evidenced by his data-point-based listing; and statistical-mathematical understanding, as evidenced by his use of a constant increase in the dotted line.

I: *Could you please elaborate on the list of the number of shoes that you have created?*

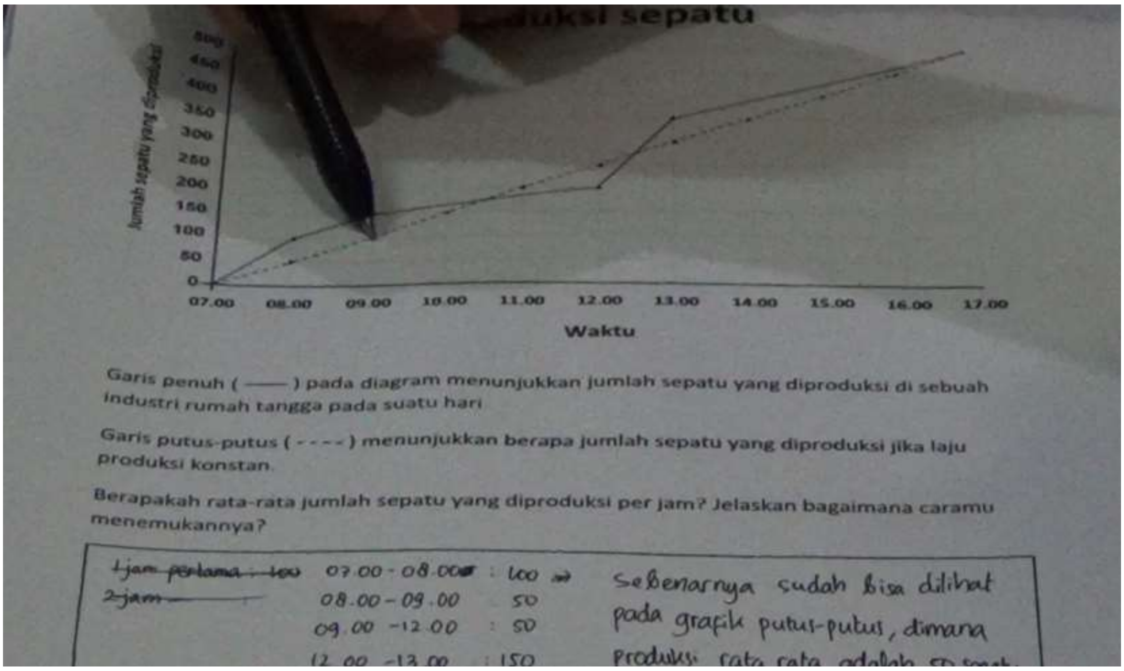
X: *This one hour* [pointing through 07.00-08.00] *produced 100 [pairs of] shoes, this one hour* [pointing through 08.00-09.00] *produced 50 [pairs of] shoes, these three hours* [pointing through 09.00-12.00] *produced 50 [pairs of] shoes, this one hour* [pointing through 12.00-13.00] *produced 150 [pairs of] shoes and these four hours* [pointing through 13.00-17.00] *produced 150 [pairs of] shoes. I then attempted to determine the mean by [initially] adding up all the shoes and also determine the total [production] time needed.*

I: *What did you do following that?*

X*: I discovered it to be the equal to the dotted line, this is* [pointing to the dotted line at 08.00] *50 [pairs of] shoes and this is* [pointing of the dotted line at 09.00] *100 [pairs of] shoes. Consequently, the interval is fifty per hour* [see Figure 7.18].

Figure 7.18

Xavier Explaining the Production Mean from the Dotted Line



Garis penuh (——) pada diagram menunjukkan jumlah sepatu yang diproduksi di sebuah industri rumah tangga pada suatu hari

Garis putus-putus (- - - -) menunjukkan berapa jumlah sepatu yang diproduksi jika laju produksi konstan.

Berapakah rata-rata jumlah sepatu yang diproduksi per jam? Jelaskan bagaimana caramu menemukannya?

~~1 jam pertama : 100~~ 07.00 - 08.00 : 100 →
~~2 jam~~ 08.00 - 09.00 50
09.00 - 12.00 : 50
12.00 - 13.00 : 150

Sebenarnya sudah bisa dilihat pada grafik putus-putus, dimana produksi rata rata adalah 50 sepat

Similarly, during the interview, Luhut (L) demonstrated an understanding of the most production by stating that production fluctuated over time and, as a result, there should be both maximum and minimum production. He can identify the most production based on the gradient in the solid line. He argued that the steepest gradient indicates the highest quantity of shoes produced.

I: *Could you please elaborate on your response?*

L: *Because the production was not constant in the line graph* [pointing through the solid line from start to end], *there must be both the highest and lowest production.*

I: *How do you know that the production was not constant?*

L: *Because this solid line is not straight* [pointing through the solid line from start to end]. *It is constant if it is straight* [pointing through the dotted line from start to end].

I: *What did you do for this [The Most Production] item? Please elaborate.*

L: *That is extremely easy. Typically, the line that is nearly straight up or nearly vertical indicates the biggest count. This is the most vertical line* [pointing through the solid line between 12.00-13.00].

I: *Does this indicate that you directly viewed the line graph without determining the number of shoes per time interval?*

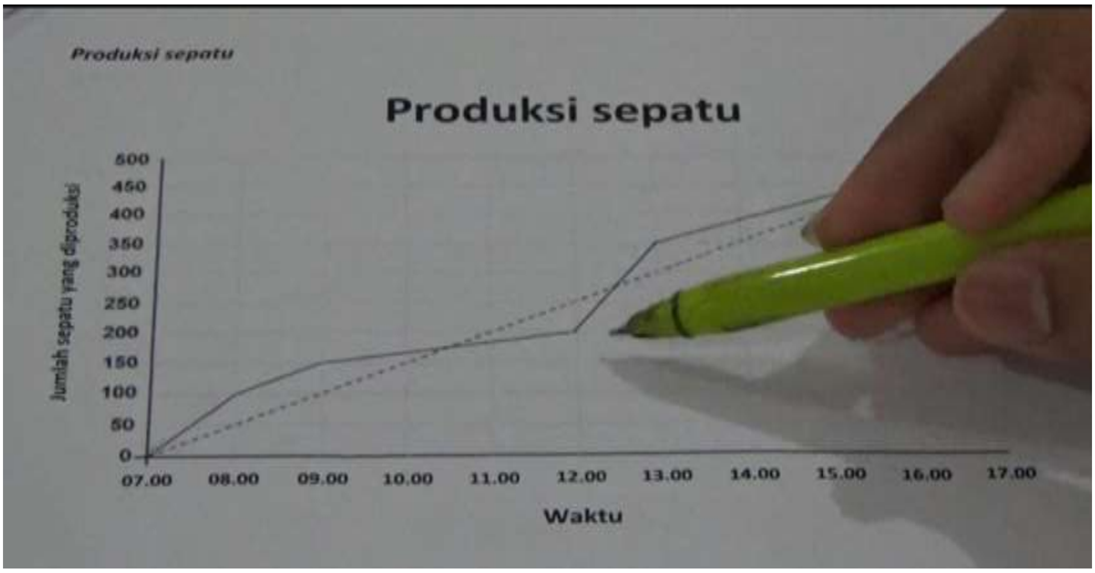
L: *I did not need to do that*

I: *Then, which time interval contains the most production?*

L: *From 12.00 to 13.00* [see Figure 7.19]

Figure 7.19

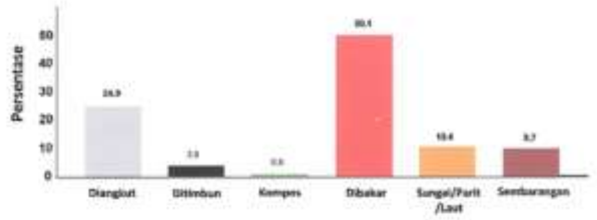
Luhut Showing the Most Production from the Solid Line



In addition to the abovementioned critical responses when interpreting statistical information, the students critically summarised data information and communicated their reaction to the information to others. Among students performing at Level 5 communicating, four of them were found in the YouTube Viewers item and eight of them were found in the Domestic Waste item. These Level 5 students demonstrate critical ability to select important components to be included in their summaries. These students showed sufficient knowledge of context, bar graph convention, statistics and mathematics ideas to include in their summaries. It was found that they compared and contrasted data including the respective figures from the bars. All these students thoroughly looked at the graph convention such as the y-axis title that was in thousands, as found in YouTube Viewers item. This response was found in Luhut (L), for example, who summarised important information from Domestic Waste item (see Figure 7.20). He compared the proper and improper waste management by contrasting the burning and composting. He also added improper waste management by including the littering. This was reflected by his responses during the interview, as follows.

Figure 7.20

Luhut's Critical Response to Domestic Waste Item



Agar temanmu paham, ringkaslah informasi penting dari diagram di atas tentang kesadarán masyarakat Indonesia akan pengelolaan sampah rumah tangga!

~~Hampir set~~ Lebih dari setengah masyarakat Indonesia membuang sampah secara ~~sembarangan~~ dibakar. Lebih dari $\frac{1}{5}$ masyarakat Indonesia membuang sampah di sungai/parit/laut/sembarangan. Dan kurang dari 1% saja yang dijadikan pupuk kompos.

~~Almost~~ More than half of Indonesian people disposed of their waste by ~~littering~~ burning. More than $\frac{1}{5}$ Indonesian people threw their waste in rivers/ditches/seas/littering, and only less than 1% was used for composting.

I: *Can you elaborate your answer, how did you know that it was more than a half of Indonesian burning their domestic waste?*

L: *Because this* [pointing bar of burning] *shows 50.1 percents.*

I: *You wrote that more than one-fifth of Indonesian throwing waste in river, ocean and littering. Can you please explain?*

L: *It is one-fifth because the total of them is 20.1 percent* [the sum of 10.4+9.7]

I: *What about less than 1 percent of Indonesian did composting?*

L: *This is to emphasize to people that only less than 1% of Indonesian did composting.*

In conclusion, the difference of Level 4 compared to Level 5-6 responses is on the complexity of their responses. Level 4 students focused interpreting and communicating data based on what they clearly saw from the graphs. They tended to do calculation and needed more language and graphical competence which could be considered as factors hindering them to advance to critical thinking. In contrast, Levels 5-6 students tried to gain more meaning from the graphs that can be interpreted and communicated. They realised that not all information needs to be computed procedurally. They tried to look on the relationships among available variables in critical ways.

7.3.2 Students' Appropriate to Critical Responses in Evaluating Data-Based Claim and Making Decisions

The appropriate and critical responses shown by Levels 4-6 students when performing data evaluation and data-based decision-making enabled them to provide relevant evidence for their arguments. Using the same evaluating items (The Employees and Mathematics Scores) and decision-making items (The 100-Metre Race and Which Motorcycle?) to reveal students' challenges, students' appropriate and critical understandings on the three knowledge components (text and context, representation and statistical-mathematical knowledge) were examined. Based on data previously presented in Table 7.3, 18 responses were found in Level 4, seven responses were found to achieve Level 5 and only one response achieved Level 6 for evaluating items. For decision-making, 12 responses were found in Level 4, 10 responses were found to achieve Level 5 and five responses achieved Level 6.

Among the students demonstrating appropriate responses, there were three students demonstrating consistency in providing relevant evidence to evaluate statistical claims and make decision based on data. Table 7.13 presents these three students and their achieved levels of four items. Their written responses were then investigated to reveal insights on their relevant evidence.

Table 7.13

Three Students Demonstrating Appropriate Evidence When Evaluating and Making Decisions

Students	Level			
	The Employees (E)	Mathematics Scores (E)	The 100-Metre Race (D)	Which Motorcycle? (D)
Hannah	Consistent non-critical	Consistent non-critical	Consistent non-critical	Consistent non-critical
Farah	Consistent non-critical	Informal	Consistent non-critical	Consistent non-critical
Qiqi	Consistent non-critical	Consistent non-critical	Critical	Consistent non-critical

Note. E is Evaluating and D is Decision-making.

For evaluating, Level 4 students successfully provided evidence to contest the existing claim made by newspaper reader (The Employees) or mathematics teacher (Mathematics Scores), as presented in Table 7.14. When challenging a claim, these students included a reasonable argument using data presented in the bar graphs through which the claim was developed. They seemed to have investigated the bar graphs' convention, particularly on the y-axis scale, x-axis label and the bar's height. For instance, the students estimated the number of employees' increase using the y-axis scale (The Employees) and made sense that the x-axis represent the score interval (Mathematics Scores). In addition, these students seemed to do investigation in relation to data set's context and the statistical content. For example, they demonstrated a good understanding of the meaning of the textual information such as a huge increase and its difference with an increase (The Employees) or used the information of the minimum passing score to classify the data in the bar graph (Mathematics Scores). It was apparent that their understanding of the three SL components served as the basis for their evidence. Their interview responses then confirmed this interpretation of their written responses.

Table 7.14

Students' Relevant Evidence in Evaluating Claims

	Written responses	Interview response
Hannah	The Employees *Based on the graph above, the increase in the number of employees from 2016 to 2017 was less than 10 [employees]. I think the huge increase should be more than 20 employees.* *So, I think it does not make sense.*	The Employees • *It can be seen [from the graph] that the difference of two points in y-axis* [indicating y-axis scale] *is 5 [employees].* • [pointing to the differences of two bars] *the difference is less than 10 [employees].*
	Mathematics Scores *There was only 1 student in class A who was under minimum passing score. There were 2 students in class B who were under minimum passing score.* *In class A, there was more than 1 student in each score interval. In class B, there was 1 student in 50-59 and 80-89 score interval.* *Class A has 2 students whose scores were in the highest interval (80-89) and 1 student whose score was in the lowest interval (0-9). Class B has 1 student whose score was in the highest interval (80-89) and 2 lowest students whose scores were in the 40-49 interval.*	Mathematics Scores • *Firstly, I looked at the bars below 50 and I found that there was 1 [student] in class A and there were 2 [students] in class B.* • *In the interval of 50-59, there were three [students]. In the interval of 60-69, there were four [students]. In the interval of 70-79, there were two [students]. In the interval of 80-89, there were two [students]. Thus, [in class A] there was no interval with only one student. [The distribution] spread quite evenly. [In contrast] there was only 1 [class B] student in this interval* [pointing 50-59] *and this* [pointing 80-89]. *Thus, [the distribution] was uneven.*
Farah	The Employees *No, because the scale in the graph is so small. The difference is only 5 people. It is assumed that the increase is from 508 to 516, which means the increase was only 8 [employees]. There was an increase, but not huge.*	The Employees • *Firstly, I looked at the bars showing huge differences. I then looked at the range* [pointing the scale in the y-axis] *which shows 505 and 510. Thus, the range is small, only 5 employees. It was not huge.*

Table 7.14 (continued)

	Written responses	Interview response
Qiqi	The Employees *It does not make sense. Because the increase of the number of employees in the graph was "not huge". Instead, it was only less than 10 [employees]. The number of employees may be high but not the increase.*	The Employees • *I predict that the increase was not significant because this is* [pointing bar of year 2017] *around 515 [employees] and this is* [pointing bar of year 2016] *around 518 [employees].*
	Mathematics Scores *The number of students passing the test:* *Class A= 3+4+2+2 = 11* *Class B = 1+5+3+1 = 10* *It shows that the number of students in class A passing the test was higher than those of class B, although the mean score of class B was higher than that of class A.*	Mathematics Scores • *It is right if the mean for class A is lower than that of class B. After that, I looked at the text informing [me] that they passed the test if their score equalled or was more than 50. Therefore, I directly looked at the number of students [in each class] who passed the test.*

Similarly, Level 4 students were successful providing supporting evidence for the choice they made. Table 7.15 summarises the interpretation of their written responses together with their confirmation during interview. They showed an ability as individual to make informed decision by examining the three SL components. For example, these students could use one measure of central tendency to choose one best runner (The 100-Metre Race) and three numerical conditions to select one motorcycle (Which Motorcycle?). It was also found that they demonstrated a good understanding of the context and graph despite their incorrect calculation of the mean (The 100-Metre Race). Some of these Level 4 students did not recognise a tax that was not included in the prices (Which Motorcycle?). Furthermore, compared to their test performance, a few Level 4 students demonstrated a higher level of understanding during the interview. Their higher understanding was reflected in, for instance, Farah's and Qiqi's interview response for Which Motorcycle item.

Table 7.15

Students' Relevant Evidence in Making Decisions

	Interpretation of Test Response	Interview response
Hannah	The 100-Metre Race *Sarah was chosen because of two reasons. First, she has decreasing time in the seven races. Second, her finishing time has a fair difference compared to previous race. It implies that Sarah was never give up and wanted to be the winner.*	The 100-Metre Race • *Firstly, I looked at the time. For Sarah, her time was getting better in the next race. Thus, if her [finishing] time was getting better or getting shorter, it could be concluded that she run faster.*
	Which Motorcycle? *Jupiter C was chosen because it met the three conditions (the distance travelled was 35,000, it was produced in 2013 and its price was 6,25 million which was cheaper than Jupiter B).*	Which Motorcycle? • *Firstly, I looked at the four motorcycles. When the year was considered, all the four met the criterion because all were produced in 2011 or after. When the price was considered, only Jupiter A did not meet the criteria, because the price was over 6.5 million. When the distance travelled was considered, the four motorcycles met the condition. Then, the options left were Jupiter B, C and D.* • *Jupiter C was then chosen as its price was reasonable and the newest in term of the year of production.*
Farah	The 100-Metre Race *Sarah was chosen after observing her finishing time over seven races. Her finishing time always decreased, meaning she run faster from race to race.*	The 100-Metre Race • *I looked at Sarah, (her time) in the first race was 15.2 second, in the second race was 15 second and [she was] getting faster over the race. For Rita, she increased the time and decreased again. For Maria, her time was always increasing.* • *The smaller the time, the faster she run* [pointing to Sarah]

Table 7.15 (continued)

	Interpretation of Test Response	Interview response
	Which Motorcycle? *Jupiter D was chosen as it met all the three conditions (it was produced in 2011, its price was the cheapest compared to others and its distance travelled was 34,800 km).*	Which Motorcycle? • *Firstly, I looked at the distance travelled which was not more than 35,000 km. Based on the table, all the four motorcycles met this criterion.* • *Secondly, the motorcycle should be produced in 2011 or after. Based on the table, all the four motorcycles also met this criterion.* • *Thirdly, the price should not be over Rp. 6,500,000. Jupiter A was not possible and moreover, there was still 2.5% tax. For Jupiter B, it actually met the criteria without tax, but if the tax was included it did not meet the criteria. This also applied to Jupiter C, and thus I chose Jupiter D.*
Qiqi	The 100-Metre Race *[Qiqi] applied add-divide formula to find the mean of each runner and found the mean of their finishing time: Sarah = 13.2 seconds, Rita= 14.9 seconds and Maria= 15.3 seconds. However, there was miscalculation in calculating the mean for Sarah and Maria. Particularly, the miscalculation occurred when adding the time for each of both runners. In the end, Sarah was chosen because she has the shortest mean time.*	The 100-Metre Race • *Because the table consists of many data, so I directly thought of finding the mean.* • *In the context of running competition, what is needed is the shortest time. So, I chose the one with the shortest time, which is Sarah.*

Table 7.15 (continued)

	Interpretation of Test Response	Interview response
	Which Motorcycle? *After comparing Jupiter B and D, Jupiter D was chosen because it met all the three conditions (the distance travelled less than 35,000, it was produced in 2011 and the price was cheaper than Jupiter B). Jupiter B was not chosen as its price after tax included did not meet the condition.*	Which Motorcycle? • *Firstly, I wrote the distance travelled which was not more than 35,000 km. I then wrote which motorcycles met this condition, namely Jupiter A, B and D. Jupiter C did not meet the condition as its distance travelled was exactly 35,000 km.* • *Secondly, I wrote which motorcycles were produced in 2011 or after. I wrote all the four motorcycles because they all met the criteria.* • *Lastly, I wrote the price before the tax was included. Jupiter B, C and D met this criterion.* • *Consequently, A was eliminated because of its price and C was eliminated because of its distance travelled. The options left were Jupiter B and D.* • *I chose Jupiter D after comparing the price of Jupiter B and D with the tax included. The tax for Jupiter B is Rp 161,250 that exceed the price [condition after added to the original price].*

By comparison, the critical understanding demonstrated by the Level 5 and 6 students differ from those of Level 4 students. Table 7.16 presents these students' critical understanding of three SL components to provide evidence when contesting a claim and supporting their choice.

Table 7.16

Students' Critical Understanding of Three SL Components When Evaluating Claims and Making Decisions Based on Data

Components	Students' critical understanding
Text and context	Students demonstrate critical understanding on the various contexts involving statistical claims or information used to make decision. The context includes education, economy, sport and trading. These contexts provide students with strategies and procedure to contest data-based claims or make a choice. For example, students understand the rule of the running competition which is the runner with the lowest time is the quickest. This understanding helps students to compare the mean of three runners along with the runners' trend. Students also know that the distance travelled is how far motorcycle was used to travel and not how far the motorcycle can travel with full tank. This understanding brings students to find the lowest distance travelled instead of the longest. In the case of evaluating claims, students recognise the provided mean and minimum passing grade in the text information to be used to compare the number of students passing mathematics test from both classes.
Representation	Students' ability to evaluate and make decision based on data presented in the representation was not limited to the types of graphs. They show critical graphical and tabular understandings. First, students successfully examine all the features of the graph and table (i.e., title, axes, headings, row and column, and legends). For example, they recognise what is shown in the row and column of the table and understand the meaning of the discontinued y-axis. Second, students successfully find what the numbers represent (e.g., by looking for the largest and smallest values in one or more categories to obtain an impression of the data). For example, students understand that the number in the table of three runners represent the finishing time recorded in seconds. Third, students successfully find the differences in the values (e.g., the changes of data over time and the comparative values of data within a category). For instance, students successfully match the data in the table and the criteria of selecting motorcycle. Fourth, students successfully identify where differences occur (e.g., using information from Step 3 to make comparisons between two or more categories or timeframes). For instance, students successfully identify the effect of tax when included. Finally, students successfully assess why those differences occurred by looking for the relationships in the data and relating them to the context. For instance, students demonstrate ability to calculate tax and compare it among four motorcycles. Students also ignore the difference of bars that almost double as justification.

Table 7.16 (continued)

Components	Students’ critical understanding
Statistical-mathematical knowledge	Students’ knowledge of statistical concept and related mathematics procedures enabled them to provide critical arguments and various approaches to challenging a claim and making decision. They, as data consumers, know the functions of statistical concepts more than the underlying calculations. They show flexibility in using average and trend in their responses. For example, students demonstrate critical questions on the consistency of claim and data presentation. By checking the consistencies of a claim, students could contest the claim using statistical concept. By observing the representation, students find evidence to justify their choice. Additionally, students demonstrate sufficient level of mathematical knowledge on rational numbers such as percentages, fractions, decimals and on number operations, especially addition and division. They use the combination of statistical ideas such as mean and the lowest record, mean and trend, total time and trend to choose the best runner (The 100-Metre Race), could show or predict how many employees increase and use it to contest the claim of a huge increase (The Employees) and explain the effect of outlier to the lower mean (Mathematics Scores). Students’ understanding of the effect of outliers to the mean helps them evaluating the claim based on the mean scores.

Among those students demonstrating critical responses, there were three students demonstrating consistency in providing critical evidence to evaluate statistical claims and make decision based on data. Table 7.17 presents these three students and their achieved levels of four items. Their written responses were then compared to reveal insights on their critical evidence. It was determined that Levels 5 and 6 students contested the existing claim and made decisions using critical and relevant evidence.

Table 7.17

Three Students Demonstrating Critical Evidence When Evaluating and Making Decision

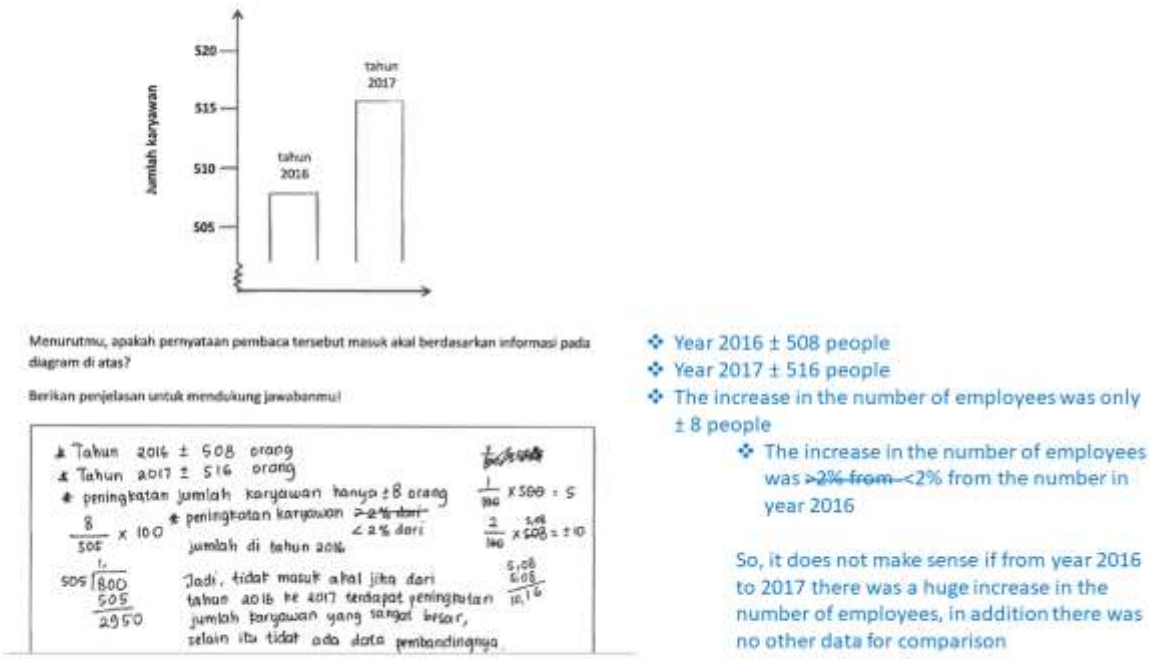
Students	Level			
	The Employees (E)	Mathematics Scores (E)	The 100-Metre Race (D)	Which Motorcycle? (D)
Vanes	Critical mathematical	Consistent non-critical	Critical mathematical	Critical mathematical
Wafiq	Critical	Critical mathematical	Critical mathematical	Critical
Xavier	Critical	Consistent non-critical	Critical mathematical	Critical

Note. E is Evaluating and D is Decision-making.

For evaluating, the critical understanding of Levels 5 and 6 students on the three components helped them to provide critical and relevant evidence when asked to contest the existing claim effectively. Levels 5 and 6 students' understanding was sufficiently enough to relate the context and data in the bar graph to reveal statistical and mathematical concepts used to refute the claim. In The Employees item, they demonstrated graphical understanding by ignoring the height of the two bars as they were influenced by the discontinued y-axis. Figure 7.21 presents the response from Vanes to exemplify the Level 6 response on The Employees item. It was interpreted that Vanes provided two types of evidence to contest the claims: the percentage of the increase and the lack of additional data as comparison.

Figure 7.21

Vanes' Critical Response to The Employees Item


Jumlah karyawan
520
515
510
505
tahun 2016
tahun 2017

Menurutmu, apakah pernyataan pembaca tersebut masuk akal berdasarkan informasi pada diagram di atas?

Berikan penjelasan untuk mendukung jawabanmu!

- Tahun 2016 ± 508 orang
- Tahun 2017 ± 516 orang
- peningkatan jumlah karyawan hanya ±8 orang
- peningkatan karyawan ~~>2% dari~~ < 2% dari jumlah di tahun 2016

$\frac{8}{505} \times 100$

$505 \overline{)800}$ 1,
505
2950

Jadi, tidak masuk akal jika dari tahun 2016 ke 2017 terdapat peningkatan jumlah karyawan yang sangat besar, selain itu tidak ada data pembandingnya

$\frac{1}{100} \times 508 = 5$

$\frac{2}{100} \times 508 = \pm 10$

5,08
5,08
10,16

- Year 2016 ± 508 people
- Year 2017 ± 516 people
- The increase in the number of employees was only ± 8 people
  - The increase in the number of employees was ~~>2% from~~ <2% from the number in year 2016

So, it does not make sense if from year 2016 to 2017 there was a huge increase in the number of employees, in addition there was no other data for comparison

During the interview, Vanes (V) elaborated her thoughts on two types of evidence she provided. She explained that she first determined the number of increase and transformed it into percentage. The percentage was calculated from dividing the number of increase (8 employees) by 505, which she assumed as the number of employees in 2016, and then multiplied by 100. She found that the increase was one point something percent or less than 2%. To ensure that her calculation was correct, she checked by finding how many 2% of 508 employees, which is the number of employees in 2016. She found that 2% of 508 is ±10 employees. Based on her calculations, she told to the interviewer (I) that the existing claim was irrelevant, as follows.

I: *How can you determine the percentage?*

V: *First of all, the difference [of the number of employees] was divided by the number of employees in 2016 and then multiplied by 100. I then concluded that the percentage is under 2%* [pointing her calculation showing 1.]. *Following* *that, I checked it and [that was right because I] found that 2% [of 508] equals ±10 [pointing her calculation on the right side].*

I: *What can you conclude from this?*

V: [Reading her writings]. *So, it does not make sense if there was a huge increase in the number of employees from 2016 to 2017. The increase was under 2%. Moreover, there was no additional data for comparison.*

I: *What do you mean by there was no data for comparison.*

V: *If we want to claim that there was a huge increase [in the number of employees], there should be lower increase as comparison.*

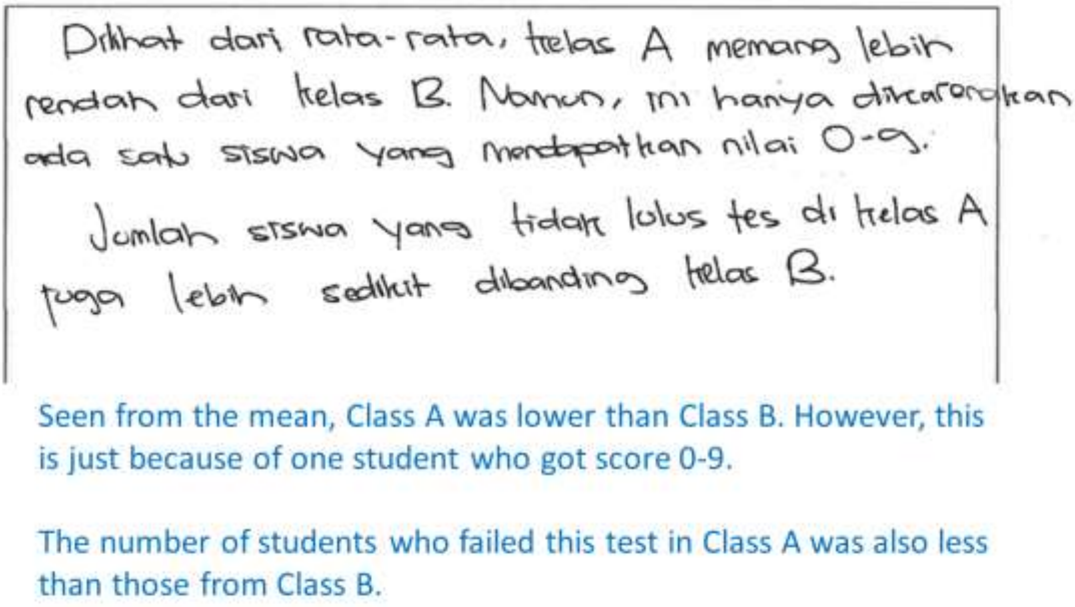
I: *In this case, what do you need as comparison?*

V: *May be the data from the previous years or the following years.*

Similarly, Levels 5 and 6 students showed a critical understanding of the context of comparing two group of students in the Mathematics Scores item. Level 5 students could utilise the information in the text about the minimum passing score to further investigate and compare the number of students in both classes passing the test. Further, Level 6 students were able to use the idea of outlier affecting the lower mean to strengthen their justification. With such critical understanding, Levels 5 and 6 students were successful to find relevant and strong data as evidence from the bar graph. It was revealed that all Level 5 and 6 responses always included the comparison of the number of students passing the test. The difference is on the complexities of statistical ideas they added to sharpen the evidence. For example, level 5 participants included the percentage of the number of students passing the test, while level 6 participants added the idea of the variance as well as outlier in class A that caused their lower mean. Figure 7.22 shows Wafiq’s responses on Mathematics Scores item.

Figure 7.22

Wafiq's Critical Response to Mathematics Scores Item



Dilihat dari rata-rata, kelas A memang lebih rendah dari kelas B. Namun, ini hanya dikarenakan ada satu siswa yang mendapatkan nilai 0-9.

Jumlah siswa yang tidak lolus tes di kelas A juga lebih sedikit dibanding kelas B.

Seen from the mean, Class A was lower than Class B. However, this is just because of one student who got score 0-9.

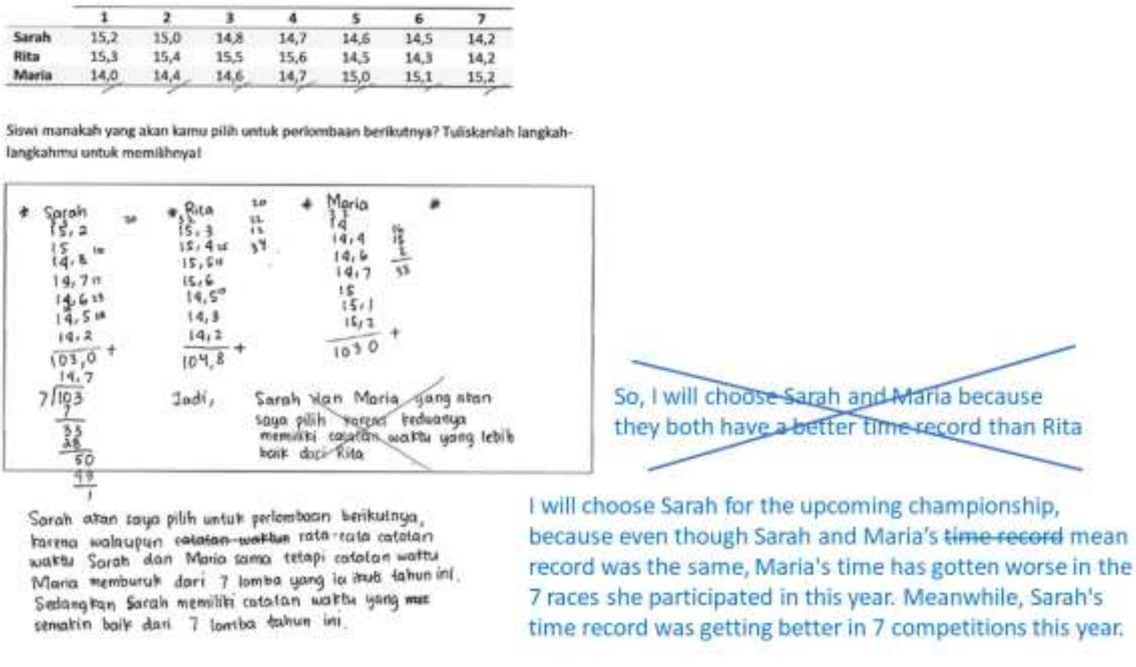
The number of students who failed this test in Class A was also less than those from Class B.

Wafiq provided two types of evidence derived from the bar graph to argue the claim, by comparing the performances of students from two classes. Wafiq demonstrated a critical understanding of the three SL components to solve this Mathematics Score item. In terms of text and context, Wafiq understood that he was asked to challenge the existing claim made by the mathematics teacher 'class B did better than class A in this test'. Using the textual information provided in the item about the mean for the two classes, he argued that although Class A had a lower mean compared to Class B, the lower mean of Class A was caused by one student who got the lowest score in the 0–9 interval. This indicated that Wafiq applied his contextual-graphical understanding when making sense of the data presented in the bar graphs. From the statistical-mathematical knowledge viewpoint, Wafiq understood about the outlier that affected the mean of Class A. This understanding of outlier is strong evidence that can be used to challenge the claim. In addition, he used the information about the minimum passing score given in the textual information to provide second evidence. He discovered that the number of Class A students who failed in the test was also smaller than that of Class B.

In the case of locating relevant evidence to support students' decision-making, the understandings of students at Levels 5 and 6 led them to make the correct choice. In The 100-Metre Race item, students' critical understanding of the context of time in running competitions resulted in them choosing the best runner correctly. They had devised the idea of comparing the mean of each runner's time, and they were indicated to choose the runner with the lowest mean. Realising that there were two runners having the same lowest mean, students at Level 6 provided additional evidence using the trends of three runners. Figure 7.23 exemplifies the work of Vanes, who used the combination of mean and trend as her method to select the best runner for the upcoming championship. Utilising the trend, a comparison was made between two runners having the same mean times across seven races.

Figure 7.23

Vanes' Critical Response to The 100-Metre Race Item



	1	2	3	4	5	6	7
Sarah	15,2	15,0	14,8	14,7	14,6	14,5	14,2
Rita	15,3	15,4	15,5	15,6	14,5	14,3	14,2
Maria	14,0	14,4	14,6	14,7	15,0	15,1	15,2

Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!

```
Sarah        Rita         Maria
15,2         15,3         14
15           15,4         14,4
14,8         15,5         14,6
14,7         15,6         14,7
14,6         14,5         15
14,5         14,3         15,1
14,2         14,2         15,2
------ +     ------ +     ------ +
103,0        104,8        103 0

   14,7
7/103
```

~~Jadi, Sarah dan Maria yang akan saya pilih karena keduanya memiliki catatan waktu yang lebih baik dari Rita~~

~~So, I will choose Sarah and Maria because they both have a better time record than Rita~~

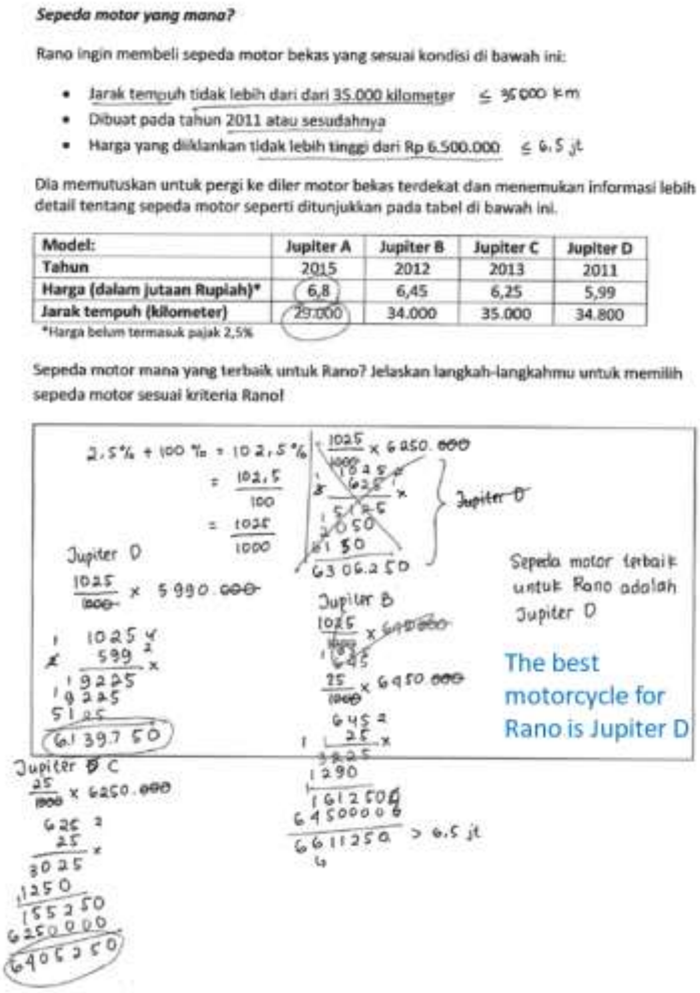
Sarah akan saya pilih untuk perlombaan berikutnya, karena walaupun ~~catatan waktu~~ rata-rata catatan waktu Sarah dan Maria sama tetapi catatan waktu Maria memburuk dari 7 lomba yang ia ~~ikuti~~ tahun ini. Sedangkan Sarah memiliki catatan waktu yang ~~me~~ semakin baik dari 7 lomba tahun ini.

I will choose Sarah for the upcoming championship, because even though Sarah and Maria's ~~time record~~ mean record was the same, Maria's time has gotten worse in the 7 races she participated in this year. Meanwhile, Sarah's time record was getting better in 7 competitions this year.

Furthermore, Figure 7.24 captures Vanes' thought process to select a motorcycle that meets the three numerical conditions. It was clear that Vanes (V) showed an understanding of the year criterion, being all the four motorcycles met the year criterion. In term of the distance travelled, Vanes showed a critical mathematical understanding of it. Her marks on the text above the table gave a clue that all motorcycles could be the options. In terms of price, she included the tax resulting in the exclusion of Jupiter B, only Jupiter C and D remained as choices. To support the interpretation of her written works, the interview of Vanes had clarified what she thought. The interview clarified her mathematical understanding on the tax and price, demonstrating her good understanding of tax when included in the price. Her understanding was perfectly captured in her written response by calculating 102.5% of each price of Jupiter B, C, and D (see Figure 7.24). Furthermore, the interview clarified Vanes' understanding of the distance travelled.

Figure 7.24

Vanes' Critical Response to Which Motorcycle? Item



Sepeda motor yang mana?

Rano ingin membeli sepeda motor bekas yang sesuai kondisi di bawah ini:

- Jarak tempuh tidak lebih dari dari 35.000 kilometer ≤ 35000 km
- Dibuat pada tahun 2011 atau sesudahnya
- Harga yang diiklankan tidak lebih tinggi dari Rp 6.500.000 ≤ 6,5 jt

Dia memutuskan untuk pergi ke diler motor bekas terdekat dan menemukan informasi lebih detail tentang sepeda motor seperti ditunjukkan pada tabel di bawah ini.

Model:	**Jupiter A**	**Jupiter B**	**Jupiter C**	**Jupiter D**
Tahun	2015	2012	2013	2011
Harga (dalam jutaan Rupiah)*	6,8	6,45	6,25	5,99
Jarak tempuh (kilometer)	29.000	34.000	35.000	34.800

*Harga belum termasuk pajak 2,5%

Sepeda motor mana yang terbaik untuk Rano? Jelaskan langkah-langkahmu untuk memilih sepeda motor sesuai kriteria Rano!

$2,5\% + 100\% = 102,5\%$

$= \frac{102,5}{100}$

$= \frac{1025}{1000}$

Jupiter D

$\frac{1025}{1000} \times 5990.000$

1025
599 ×
9225
9225
5125
6139750

$\frac{1025}{1000} \times 6250.000$

~~1025
625 ×
5125
2050
6150
6306250~~

~~Jupiter D~~

Jupiter B

~~$\frac{1025}{1000} \times 6450000$~~

645

$\frac{25}{1000} \times 6450.000$

645
25 ×
3225
1290
161250
6450000
6611250 > 6,5 jt

Jupiter ~~D~~ C

$\frac{25}{1000} \times 6250.000$

625
25 ×
3025
1250
15525
6250000
6405250

Sepeda motor terbaik untuk Rano adalah Jupiter D

The best motorcycle for Rano is Jupiter D

I: *What did you do to help Rano in selecting a motorcycle?*

V: *Initially, I examined Jupiter A. The Year was 2015, so it satisfied the year criteria [produced on 2011 or later], but the price was 6.8 million [Rupiahs], so it did not meet Rano's criteria.*

I: *What about Jupiter B, Jupiter C and Jupiter D?*

V: *They all [pointing to Jupiter B, C and D on the table] met [Rano's] criteria, [but] the price excluded tax.*

I: *Did you understand what 'excluding tax' means? and what impact did it have?*

V: *If the tax was not included, then there would be an increase from the tax. Then, the price would exceed the price listed in the table.*

I: *What did you do once you realised that the tax was excluded?*

V: *First of all, I examined the price and the tax. The [new] price for Jupiter D [including tax] was 6,139,750 Rupiahs [pointing her calculation for Jupiter D]. The [new] price for Jupiter C [including tax] was 6,405,250 Rupiahs. The [new] price for Jupiter B [including tax] was 6,611,250 Rupiahs which did not fulfil Rano's criteria.*

I: *The remaining options were Jupiter C and D. How could you select?*

V: *Both Jupiter C and D met Rano's price and year criteria, but I was little confused. Then, I compared the first criterion about the distance travelled that should not be more than 35,000 [kilometres]. This [distance travelled for Jupiter C] equals 35,000 [kilometres] and then I chose the one lesser [of Jupiter D].*

I: *What is it 'the distance travelled'?*

V: *The distance that a motorcycle has travelled.*

I: *Is there any part of a motorcycle that shows it?*

V: *In speedometer.*

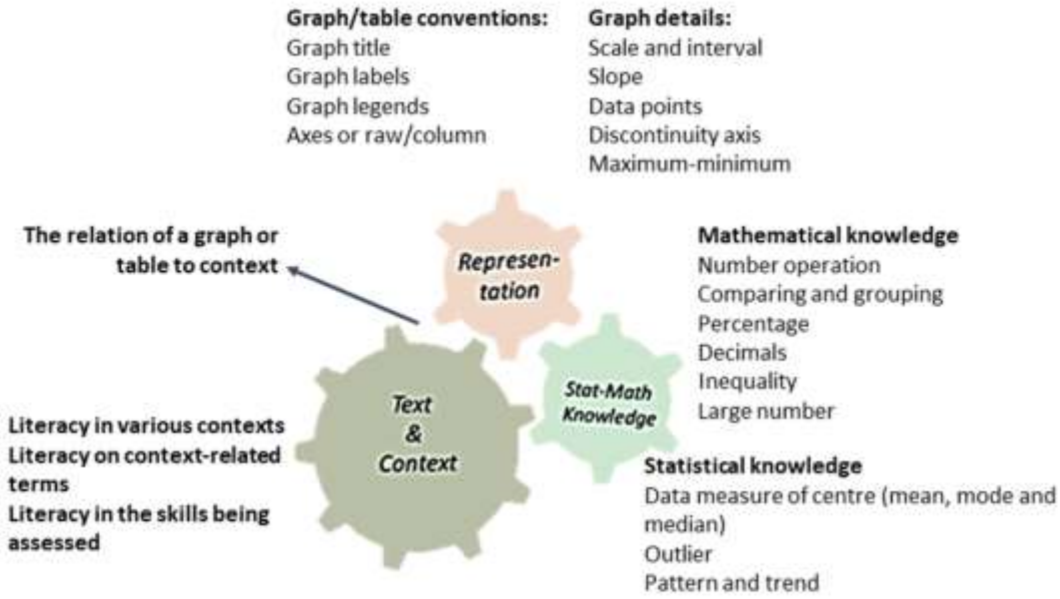
In summary, there is noticeable distinction between Level 4 and Level 5–6 responses concerning the complexity of the evidence presented. Although most of the evidence presented by Level 4 students is appropriate, occasionally they ignored certain important details from the graph and table. It was also discovered that they occasionally made a small calculation error. These factors might be seen as impediments to their progression to critical thinking. In contrast, students in Levels 5–6 provided critical and relevant evidence to contest a claim or support a choice. They realised that the basis for contesting a claim should be its own inconsistency. These students also recognised that a decision need to be supported by solid evidence.

7.3.3 Summary of Students' Appropriate and Critical Understandings

Given that the three components are interrelated, students' solid comprehension of these components enabled them to provide critical responses when interpreting, communicating, evaluating and making decisions. Students at Levels 5–6 demonstrated a critical understanding of data in context. Their contextual understanding enhanced their ability to comprehend data presented in graphs and tables, and vice versa. As evidenced by their responses, they differ from the majority of students in their ability to select statistical concepts and execute complex mathematical calculation, including Level 4 response. As a result, these students demonstrated the ability to extract qualitative meaning from data when interpreting statistical information for themselves, highlighted trends, made comparisons, showed relationships and included the most significant data in their summary when communicating their comprehension to others. These students also provided critical arguments and various approaches when evaluating a claim and provided statistical reasoning and problem-solving when asked to make decision. Figure 7.25 shows the specifics of students' understandings of the three components and their interrelationship, causing them to critically respond to statistical information based on the assessed skill.

Figure 7.25

Students' Understandings of the Three Components of SL



7.4 Discussion

This section presents two discussion topics that emerged from the qualitative findings. Section 7.4.1 discusses the interrelations of the three components of SL and how they affect students' challenges and understanding while making sense of data-based information. Section 7.4.2 discusses the need for an SL environment with an emphasis on the role of all stakeholders in improving the SL of Indonesian students.

7.4.1 Interrelationships of the Three Components of SL

One of the primary objectives of the qualitative component of the present study was to gain a better insight of the challenges students faced when solving data-based problems. Concerning the quantitative part, the results revealed that the challenges the lower group students (Level 1–3) had in comprehending the three SL components are interrelated. This means students' challenges on one component affected and was affected by their challenges on other components. This interrelationship among contributing components in relation to students' SL was previously clarified by Watson (2006). As evidenced by the level distribution of student performance on the three components of SL, no dominant component was identified as influencing students' responses. However, the qualitative analysis found that the interrelation between students' contextual-graphical challenges might contribute more to students' overall challenges. This emphasises that attaining a certain SL level is dependent not only on computational skills, but also on context knowledge and the capacity to interpret and evaluate statistical data (Jureckova & Csachova, 2020).

The qualitative findings regarding the interrelationship between students' contextual-graphical challenges provided empirical evidence to Gal's (2002) and Watson's (2006) theory on these contributing components. This finding emphasised that data must be interpreted in relation to its context (OECD, 2018) in order to be useful (Franklin & Bargagliotti, 2020). The findings of this study indicated that when students relied on their textual and contextual misunderstandings to interpret the data in graphs and tables, it becomes distracting. When context is lost, a statement loses all meaning and becomes deceptive. For example, in The Most Production item, students translated the most production as the greatest number of shoes produced. Due to this misunderstanding, they searched for the greatest number on the y-axis of the line graph. This contextual misunderstanding was also observed in The 100-Metre Race item, where the students failed to comprehend the rules of the running competition that differ from the other games. In the running competition, the quickest are those with the lowest time, not those with the highest time. Due to this misunderstanding, students used irrelevant evidence to choose the one out of three runners as they focus on the highest number in the table instead of the lowest number to find the winner in each race.

Moreover, graphical and tabular conventions become an additional source of students' challenges. Within the category of graph and table comprehension errors, the analysis revealed that students' errors were attributable to their lack of understanding of graphs features. Students tended to use and calculate the numbers displayed in graphs or table without considering their relevance to solving the task. This finding is consistent with Bursal and Yetiş (2020) and Sharma et al. (2011), who discovered that students struggled with questions requiring higher graph competence. Data revealed that students encountered challenges to understand the graph labels, graph legends, axes, raw-column and discontinued axis. In The Employees item, for example, students did not recognise the discontinued y-axis to show the number of employees. As a result, they focused on the bars' heights to determine whether there is a huge increase in the number of employees. Due to the bars' height visually displayed what looked to be a big difference, students eventually used this visual difference as justification for the huge increase. In light of this finding, the ability of students to interpret quantitative information presented in tables and graphs is an essential aspect of their SL that requires explicit instruction (Kemp & Kissane, 2010).

The five-step framework proposed by Kemp and Kissane (2010) seems promising to foster critical understandings of tabular and graphical data among students. This five-step framework successfully assisted students at the primary through tertiary level to interpret tables and graphs. It is then anticipated that it can be used to facilitate students' graph reading from idiosyncratic to critical mathematical levels, because the interrelation of SL components is addressed throughout the five steps. The fifth step in Kemp and Kissane's framework requires students to be able to identify the reasons for the data's relationships and relate them to the context. Certainly, in order to achieve this level, students must first: examine all the graph's and table's conventions (i.e., title, axes, headings, legends, footnotes and source) to discover the context; determine what the numbers represent to gain an understanding of the data; determine the differences in the values (e.g., the differences between the data in rows or columns, the changes of data over time and the comparative values of data within a category);

and determine where differences occur. All these competencies should be sufficient for them to attain critical mathematical level when responding to statistical information.

Lastly, exposing the interrelationships of SL components contributes qualitatively to the limited assessment studies that classify students' responses into three components and into six hierarchical levels. This study supplements Koparan and Güven (2015), who devised only descriptors for data representation across Watson's (2006) six hierarchical levels. This study also compares its findings to those of Yotongyos et al. (2015), who classified students' levels of knowledge of contributing components into high, moderate, and low categories based on their mean score on a 7-point Likert scale. However, not all of Gal and Watson's components were included in this study. Therefore, additional research is required to determine the effect of other SL components, such as task format and task motivation, on students' SL.

7.4.2 Call for an SL Environment

Given that only a small percentage of high school students showed evidence of critical thinking, as discovered in PISA and this study, a movement to establish an SL environment in Indonesia needs to be initiated from both inside and outside of schools. The inability of most Year 12 students to provide critical responses further emphasizes the need for SL in the final years of schooling as they are closer to adults in age and are expected to become statistically literate citizens in the future. Outside school, they are constantly exposed to data-based information (e.g., Büscher, 2022a; da Silva et al., 2021; Franklin, 2021; Franklin & Bargagliotti, 2020; Gonda et al., 2022; Suarez-Alvarez, 2021; Watson & Callingham, 2020; West & Bergstrom, 2020); however, statistics education in high school is often taught in mathematics classrooms with relatively limited time allocation (Büscher, 2022a; Zieffler et al., 2018). In light of this fact, it is the responsible of both education stakeholders and public to improve students' SL (Marchy & Juandi, 2023; Ferligoj, 2015). Teachers and other education stakeholders are responsible with the SL environment within schools; while the general public and data-related organisations, such as statistical agencies and news media, are responsible with SL outside of schools.

Inside of schools, teachers must be able to reflect on whether their statistics instruction and assessments are consistent with the curriculum and textbooks. It is presumed that the teaching thus far may not align with textbooks, preventing students from attaining the curriculum goal. Teaching SL differs from teaching students with mathematical procedures and, as implied in the K13 goal (Kemdikbud, 2012); teaching statistics should enable students to implement their statistical knowledge outside of school. This curriculum was specifically implemented to improve the quality of Indonesian education, particularly to improve students' performance on international tests, such as PISA and TIMSS (Zulkardi & Putri, 2019). This effort was reflected in mathematics textbooks that were designed to provide students with opportunities to become statistically literate by incorporating 5L activities addressing a variety of skills needed by data consumers (see Chapter 3). Theoretically, statistics instruction should refer to these mathematics textbooks, and students should develop into critical consumers of statistics. However, it appears that the impact of this curriculum on students' SL has not been particularly significant since its official implementation in 2013, as evidenced by this study's findings as well as UN results. Notably, many students still encountered challenges in responding to data-based information. During the years 2012-2019, students' performances in UN were also poor although almost all UN items only required them to perform basic interpretation.

To enhance its education, Indonesia implemented a new national curriculum and assessment, and there is a greater need for statistically literate teachers to implement it in the classroom. The Indonesian new curriculum, called *Kurikulum Merdeka*, focuses on literacy and numeracy, especially for primary schools; while the national examination, called the *Asesmen Nasional*, began in 2021 (The Regulation of the Minister of National Education, Culture, Research and Technology, 2021; Kharismawati, 2022). This curriculum, which was implemented nationwide in 2023, is an alternative to K13 with a pilot program in Years 1, 4, 7 and 10. The mathematics textbooks recommended for this curriculum were officially published for all grade levels in 2021-2022. Statistics topics are included in all grade levels except in Years 6, 8 and 12. New statistics topics are taught to high school students, including survey sample in Year 9 (see Tim Gakko Tosho, 2022) and scatter plots, linear regression and correlation analysis in Year 11 (see Susanto, et al., 2021). These changes indicate that *Kurikulum Merdeka* and textbooks provide increased support for high school students' SL. These changes also indicate that teachers with SL are urgently needed, as the teaching practice is highly dependent on teachers' competence. Only teachers with SL are capable of designing effective SL instruction and assessment. These teachers could explicitly teach high school students the skills necessary to comprehend information (Budgett & Rose, 2017; Koga, 2022a) as well as develop their analytical and critical thinking related to SL (Jureckova & Csachova, 2020).

In accordance with previous research (e.g., Bursal & Yettiş, 2020), this study's finding suggests that graphical competence should also be incorporated into the school curriculum. Reading bar graphs in newspapers, infographics on television and line diagrams displaying our heart rate data has become a daily occurrence (Ludewig et al., 2020). This indicates that the evaluation and interpretation of any form of data presented in tables or graphs is not exclusive to mathematics (Jureckova & Csachova, 2020). Moreover, it is frequently reported that students' levels in interpreting distinct graph types vary (Bursal & Yetiş, 2020). Internationally, 60% of Year 8 students participating in TIMSS could only read a single value from a line graph, and only 29% could ascertain the average from line graph (Ludewig et al., 2020). Nationally, only 19% of Indonesian students were able to identify the average from a line graph (see Section 2.5). One of this study's finding also showed that large number of students failed to correctly interpret line graph (see Section 6.1.2). These findings emphasise that the line graph is more challenging than other types of graphs (Arteaga et al., 2021). Investigating the factors that influence students' ability to extract and use information from graphs becomes crucial (Ludewig et al., 2020), and this study provides teachers with the findings of these determining factors (see Sections 7.2 and 7.3).

To complement the duties of teachers within the classroom, it is the responsibility of officials to provide accurate information to the public (Engledowl & Weiland, 2021). In today's complex information society, a comprehension of statistical information is crucial for both personal and professional success (Klein et al., 2016). In most countries in the world, newspaper coverage reflects the statistics offered to the general public and country’s consumption of statistical information (Klein et al., 2016; Tarran, 2017). In the context of Indonesia, it is presumed that students and the general public interact with social media news sources more than newspapers. Data presentations are frequently used in social media to facilitate readers’ comprehension. Officials such as the Central Agency on Statistics of Indonesia and survey agencies could therefore utilise social media to disseminate accurate statistical information to the general public, including those in the education sector. In addition, Indonesia will hold presidential, gubernatorial and local leaders’ elections in 2024; data polling will be ubiquitous, and data-based arguments will follow. Survey agencies are required to present their survey in an infographic that is simple to comprehend. They may need to conduct seminars on statistics for students so that these students can comprehend statistical results published in media (Ferligoj, 2015). This is to reduce and prevent students' exposure to pervasive misrepresentations and misinterpretations.

In conclusion, with the support from several actors inside and outside of schools, the SL environment will become a reality. Developing students' SL is undoubtedly a process, as SL develops gradually and at varying rates for each student (Jureckova & Csachova, 2020).

Due to the importance of SL in both personal and professional contexts (Klein et al., 2016; Marchy & Juandi, 2023), the demand for a greater emphasis on SL in school curricula intends to develop active and critical citizens. The key to transforming the curriculum and mathematics textbooks into SL-based instruction, from the perspective of education stakeholders, is the presence of teachers who are statistically literate. In addition, officials are responsible for producing accurate data and communicating it to the general public. Students are expected to demonstrate critical thinking, as even high-quality data can mislead readers (Wild, 2017). This study provided an assessment framework that can also be used as instruction, focusing on the skills needed by data consumers and components contributing students' SL. Regardless of the curriculum, the essence of SL must be at the centre of statistics education. In the near future, it is anticipated that Indonesian students will attain SL at sufficient level.

7.5 Chapter Summary

This chapter presents the answers for the qualitative component of this study, from which generalisations can be drawn. Concerning the third research question about students' challenges in the three SL components, the highest proportion of students in the lower group (Level 3 students) struggled more with interrelationship between contextual and graphical understanding than with statistical-mathematical knowledge. Students at Level 3 demonstrated an initial understanding of statistical concepts such as mean and its arithmetic procedure, but their comprehension was dependent on their familiarity with contexts and the forms of representation. Consequently, they failed to provide appropriate responses to statistical information. Given these findings, teaching statistics and administering SL assessments should focus on enhancing students' understanding of the three SL components and their ability to formulate responses based on the assessed skills.

To answer the fourth research question, students' appropriate and critical understandings of three SL components were revealed. Students at Levels 4–6 demonstrated contextual data comprehension. Their ability to comprehend data presented in graphs and tables was enhanced by their contextual knowledge, and vice versa. As evidenced by their responses, Levels 5–6 students’ ability to select statistical concepts and implement complex mathematical calculations differed from that of the majority of students, including Level 4 students. Because of their critical understanding of three SL components, these students responded critically to statistical information. In light of these findings, the SL environment could maintain students' capacity to be both statistically literate students and informed citizens.

Based on those findings, the interrelation among SL components and the need for an SL environment were discussed. It was discovered that the students’ challenges can be more influenced by how well they comprehend the interrelation between the context and graphical representation. This highlights that achieving a certain SL level requires not only computational skills, but also context knowledge and the capacity to make sense of graphical and tabular conventions. In order to critically comprehend tables and graphs, students require assistance from teachers. More importantly, statistically literate teachers and those from outside of schools—such as official government, the Central Agency on Statistics of Indonesia and survey agencies—must collaborate to encourage the creation of SL environment for students.

The following chapter provides a summary of this study. It begins with the study’s findings and moves on to the study’s contributions and limitations. The implications of this study for future research and teaching instruction are then elaborated.

Chapter 8: Study Summary

8.1 Summary of the Findings

This study was a cross-sectional study employing a quantitative-qualitative design. The quantitative findings presented in Chapter 6 revealed Indonesian Year 9 and 12 students' SL levels as well as differences in the SL of Indonesian students from various cohorts. The qualitative findings presented in Chapter 7 revealed students' challenges with and understandings of the three SL components when responding to statistical information. These findings were summarised as follows.

What levels of SL do Indonesian high school students possess?

To answer this first research question, a double coding procedure was employed to determine the students' level distribution across the six hierarchical levels for the Year 9 and Year 12 students. The results include the students' SL levels, skill levels, component levels and item component levels.

In terms of students' SL levels, the highest percentage of Year 9 and 12 students performed in Level 4 (consistent non-critical). This indicates that the highest number of students were appropriately but not critically using statistical thinking in their responses. However, there was a relatively big number of Year 9 students who performed in Level 3 (inconsistent), which means they were inconsistently using statistical thinking in their responses.

In terms of students' skill levels, the highest percentage of students were distributed across Levels 3 and 4 in almost all skills for both grade levels except in the interpreting skill for Year 9 students and the communicating skill for Year 12 students. In the interpreting skill, the highest number of Year 9 students used informal thinking (Level 2). Comparatively, there were no Year 12 students who used limited statistical thinking in communicating their responses, with the highest number of them demonstrating consistent non-critical thinking (Level 4).

In terms of component level, the highest number of Year 9 and Year 12 students was at Level 4, and Level 6 remained unachieved by both grades. In terms of item component level, students in both grades generally found all three components are closely interrelated, meaning finding challenges in one component influences challenges in the other components. The three components for a single item were mostly at the same level of difficulty. However, there was exception for Year 9 students, in which they found representation was difficult to decode.

Finally, Year 12 students outperformed Year 9 students in overall SL, the four skills and three components. This indicates the development in the students' SL across grades—though statistical analysis is needed to confirm this finding.

Are there any significant differences in Indonesian high school students' SL based on their demographic backgrounds (i.e., grade level, gender, school type, school status or city of origin)?

A series of Mann-Whitney *U* tests were performed to answer the second research question. The Mann-Whitney *U* tests were used to investigate whether there were differences in students' SL from different backgrounds, including grade level (Year 9 or Year 12), gender (boy or girl), school type (MoRA or MoEC-RT), school status (public or private) or city of origin (Jombang or Surabaya). This second research question suggested that all the statistical data analyses should be run within the SL assessment framework, in which statistical analyses were conducted in relation to four response skills and three components.

Several conclusions can be drawn from the results of the Mann-Whitney U tests. For students' SL levels, the analyses results revealed there were statistically significant differences based on grade level (in favour of Year 12). However, there were no statistically significant differences in SL levels based on the other variables (student gender, school type, school status and city of origin). Looking further into the four response skills, statistically significant differences were only found based on grade levels. Particularly, Year 12 students performed statistically higher in every skill except for interpreting. Consequently, further analysis was conducted to find out why there were no statistically significant differences in the level of interpretation skill between Year 12 and Year 9 students. It was most likely a result of the students' challenges in interpreting data in the line graph with the context embedded.

How do the challenges students encounter in comprehending the three components of SL affect their abilities to respond to statistical information?

To reveal the students' challenges with the three SL components when responding to data-based information, a CCM was conducted on 24 students' written response and their interview. However, the analysis was focused on Level 3 students' challenges with the three components of SL when interpreting, communicating, evaluating and making decision. Because the three components are interrelated, students' challenges in one of the three components affect their challenges in the other two. The lower group of students, particularly Level 3 students, faced challenges in making sense of the context embedded in the representation, the data in the representation and the statistical-mathematical knowledge that arises from the representation. As a result, when they were asked to interpret and communicate data-based information, the lower group students were only able to read the value from the representation and were unable to choose the most significant features to report. If they performed the calculation, the result was procedurally correct but contextually incorrect and the contextual explanation and numeric information used was inappropriate. When evaluating the claim and making decisions, students hardly found relevant evidence from the representation to challenge the claim or to support their choice. Their irrelevant evidence caused them to support the existing claim and make wrong decision. Considering that many students found challenges and lacked critical thinking, an investigation was conducted on students' appropriate and critical understandings in order to provide insight into their cognitive processes.

How do students' understandings of the three components of SL influence their abilities to respond to statistical information?

To reveal the students' understandings of the three SL components when responding to data-based information, a CCM was conducted on 24 students' written response and their interview. The analysis was focused on 1) Level 4 students' appropriate understanding and the challenges they faced in transitioning to critical understanding and 2) Levels 5–6 students' critical understanding of the three components of SL when interpreting, communicating, evaluating and making decisions. Students at Level 4 concentrated more on the computation, approaching data-based problems as they solved mathematical problems. They occasionally ignored certain important details from the graph and table and made a small calculation error. Furthermore, they needed more language and graphical competence which could be considered as factors hindering them to advance to critical thinking. In contrast, students at Levels 5–6 demonstrated a critical understanding of data in context. Their solid contextual understanding enhanced their ability to comprehend data presented in graphs and tables, and vice versa. As evidenced by their responses, they differed from the majority of students, including Level 4 students, in their ability to select statistical concepts and execute complex mathematical calculation. As a result, these students demonstrated the ability to extract qualitative meaning from data when interpreting statistical information for themselves and highlighted trends, made comparisons, showed relationships and included the most significant data in their summary when communicating their understanding to others.

These students also provided critical arguments and various approaches when evaluating a claim, and statistical reasoning and problem-solving when asked to make decision.

8.2 Study Contributions

By assessing SL in Indonesia, which is contextually and culturally different to Western countries, this study makes three substantial contributions to the field of SL assessments. This study contributes to the conceptual, methodological and practical aspects of understanding SL.

This study contributes conceptually to the field of SL assessment by proposing an innovative assessment framework (see Figure 2.4 in Chapter 2) to characterise students' SL levels from the data consumer perspective. This framework involves four SL skills as the construct (interpreting, communicating, evaluating and decision-making), which were developed based on theories (e.g., Budgett & Pfannkuch, 2010; Callingham & Watson, 2017; Gal, 2002; Sharma et al., 2012; Wallman, 1993; Watson, 2006; Watson & Callingham, 2003). Three components (sub-constructs) were added to the definition of SL—text and context, representation and statistical-mathematical knowledge—contributing to the novelty of this study.

Regarding the methodological aspects, this study's assessment framework provides a thorough process to determine students' SL through four SL skills and three SL components as well as identify students' challenges and critical understandings. The combination of quantitative and qualitative investigations in this cross-sectional study provided insights into students' SL and in-depth investigation into students' thought processes. From a quantitative standpoint, the assessment framework used in this study enabled a comprehensive and intricate investigation of the students' level in each of three SL components and four SL skills. The component level assisted in determining the order of component difficulty that contributed to the students' low and high SL level. From a qualitative standpoint, the assessment framework enabled an investigation into students’ challenges and understandings when making sense of data-based problems. Students’ challenges and understandings were related to the three SL components. This method of investigating SL has not been widely covered in previous assessment frameworks, except for a few studies focusing on a particular skill (such as interpreting) or a particular component (such as graphing).

In addition, this study extends a cross-sectional study employed in the field of statistics education. This study provides at least two significant advantages of employing a cross-sectional design: obtaining data from different cohorts, especially from different age groups, in a short period of time; and allowing the identification of differences in the students’ SL between groups. Considering the characteristics of Indonesian education system, this study can be used to rapidly investigate the SL of students from diverse backgrounds including grade level, gender, school type, school status and regional locations. For example, the SL levels of students from non-adjacent grade levels can be used to observe the potential progress in the students’ SL. Investigating the SL of students from different gender, for instance, can also be used to promote equity in education.

For the practical aspects, this cross-sectional assessment study in a developing country expands on prior studies in this field, which were mostly conducted in Western and developed countries. To date, studies on characterising high school students’ SL levels have mainly been conducted in Western or developed countries: Australia and New Zealand (e.g., Callingham & Watson, 2017; Callingham et al., 2016; Merriman, 2006; Pfannkuch, 2005; Watson & Callingham, 2014), Europe (e.g., Olande, 2014), America (e.g., Groth, 2014) and Asia (Aoyama, 2007; Koparan & Güven, 2015). Only a few studies have been conducted in developing countries, such as Fiji and Thailand (Langrall et al., 2011; Sharma, 2006; Yotongyos et al., 2015). Due to the different characteristics of countries’ education systems, the term ‘statistical literacy’ might be interpreted differently among education stakeholders.

Therefore, characterising students' SL levels in developing countries could contribute to this field and enrich the literature.

8.3 Limitations of the Study

Despite the substantial contributions, two limitations are evident in this study. These limitations were found in relation to the test items and the study participants. Limitations relating to the test items included the number of items used to measure each SL skill and the type of representation used to assess each skill; while limitations relating to the study participants included the number of participants and the region where this study was conducted.

Although the items in this study represented current examples of SL assessment, the number of items used to assess each of the four skills may not be sufficient. Each SL skill was measured by only two items. Moreover, the current investigation presented SL items that contain graphic element and each skill included only one out of three data representations (i.e., table, bar graph, or line graph). Consequently, the analysis was limited to data that contain a graphicacy component and students' skill level may be closely related to one type of representation. Further, it is acknowledged that SL items do not necessarily require a graphic. Elsewhere, Gal and Geiger (2022) noted that new demand on SL goes beyond those elements contained in the eight questions presented in this study. Gal and Geiger (2022) identified nine separate categories of information that is typically included in items that require the coding of SL. In their analysis, not all SL items required the interpretation of a graphic. In light of those limitations, non-graphic items need to be considered.

Although the results of this study can still be generalised to students in a similar context, limitations relating to study participants must also be acknowledged. First, the study participants were recruited from only two cities within one Indonesian province using stratified, purposive and convenience sampling. In actuality, the province was East Java, one of the provinces on Java Island, where high school students performed better than those in other provinces, as evidenced by the UN (Indonesian National Examination) results. Second, the number of participants was less than 100. This number was selected considering the double-coding process' burden. In light of those limitations, the students of Indonesia would be better represented by a wider region made up of students from various islands. In addition, a larger sample size would provide a more reliable measure of the Indonesian students' SL.

8.4 Implications for Future Research

The findings and limitations of this study provided avenues for future research in the field of SL assessment studies. Future research could modify the framework, include a wider range of items and involve a wider range of participants.

In this study, the assessment framework was piloted and used for data collection to measure students' SL. This assessment framework can be used to monitor and measure students' SL from the data consumers' perspective. However, only three components, one of which is the representation, are used to assess four SL skills. Regarding these components, future studies may make modifications to their application. First, future studies could retain the same three components and include additional items that assess each skill using distinct forms of representation. Second, future research could incorporate additional components from Watson's (2006) theory, such as task format and task motivation. Thirdly, future research could omit representation as a component in order to ensure the framework's applicability to non-graphics tasks, in which case, a replacement component may be a literacy measure. In addition, future studies could include actual statistical data and information from the news media disseminated online by government officials that corresponds to the revised framework.

In terms of participants, this study already included students from diverse demographic backgrounds. The differences in students' SL from various demographic backgrounds (grade level, gender, school type, school status and origin city) were identified. First, in terms of grade-level differences, this study was consistent with previous research indicating that students' SL increases as they progress through the grades. Nevertheless, only students from two non-adjacent grade levels participated. Further research could include more students from various grade levels, for instance, ranging from elementary to high school. Second, in terms of gender differences, this study’s finding was consistent with the PISA results; there were no significant gender differences in the SL of Indonesian students. This study did not, however, investigate the reasons for this absence of gender differences in the SL of Indonesian students. This requires further investigation to shed light on it. Third, no other study has compared the differences between student’ SL based on their school type, school status and city of origin. Additional research is required to confirm the findings of this study by involving more students from each group. In addition to the aforementioned, the number of students participating in this study was below 100, and they were selected from only two cities in the province of East Java. Future research could enhance the number of students in order to adequately represent Indonesia. If future research would be conducted outside of Indonesia, the researchers could consider the country’s demographic composition.

8.5 Implications for Teaching

The result of this study has implications to both the field of assessments and pedagogy in SL, particularly for Indonesian high school students. The framework developed in this study contributed to the field of students’ SL assessment from the data consumer perspective. The framework enabled a complex investigation into students’ SL levels. Particularly, it could reveal in which SL skill students lack statistical critical thinking. Further, the framework enabled a complex investigation into SL components influencing students’ responses. As a result, it became clear what needs to be focused on, both in teaching and assessment. From a learning perspective, the assessment framework can be used to target individuals learning progress (formatively) and align instruction to the specific skill. For instance, when it is discovered that students encounter challenges in comprehending graphical data to evaluate data-based claim, teachers can focus their instruction on both graphs and evaluation skills with the goal of having students demonstrate critical understanding of the graph to contest the claim.

The existence of grade difference and the absence of students in the critical mathematical level suggested pedagogical implications. The statistics instruction in Indonesian high schools should be more focused on critical statistical thinking. This is supported by the curriculum suggesting providing opportunities for students to develop critical thinking. In addition, students may rarely be confronted with statistical information to respond to. Thus, teachers may incorporate statistical information in their statistics lessons, including statistical reports appearing in online or printed media. As a result, teachers could facilitate students to critically respond (interpret, communicate, evaluate and make decisions) to such information. This would be useful experience for students to practice their critical thinking in preparation for encountering statistical information outside of school. The students' stages in performing critical responses when interpreting, communicating, evaluating and making decision could also be observed to move them to higher levels.

In summary, the ability to respond to statistical information is becoming increasingly important for all students to become informed and well-educated citizens. Moreover, the amount of data-based information that students must respond have increased as a result of technology development. Having a robust SL assessment framework such as the one developed and employed in this study allows teachers to assess students' SL in sophisticated ways and facilitate students' learning to think and reason statistically. Thus, this study helps to sustain statistics education research in the future, notably in the area of SL assessment.

8.6 Reflection

The PhD is the peak of academic achievements and the highest degree awarded to students. Interestingly, each PhD candidate’s journey is unique. Based on his experience, the researcher is pleased to share that he successfully completed this PhD with a thesis and one publication on SL. This suggests that he has gained significant academic knowledge in the field of SL. He therefore wanted to take a moment to recount his experience of working on his PhD at the University of Canberra. Below are some brief reflections on his PhD journey and what the study findings mean for his future professional development.

During his study, he was always thinking back on how much his academic writing and thinking had improved. At first, he was just a student with a three-page research proposal on SL, yet his ideas were too broad. His supervisors therefore instructed him to specify his research focus to either SL teaching and learning or SL assessment. Even though it necessitated complex rethinking, he eventually decided on SL assessment as his research focus in the hopes of improving himself as a good assessment designer. This was the first crucial choice he had to make during his candidature that would have an impact on the entire PhD process. After surviving for more than five years, he was finally able to complete his PhD. This lengthy process paid off when, at last, he submitted his lengthy thesis (approaching 390 pages). More importantly, he felt his writing and thinking had improved after reading encouraging remarks from thesis examiners. Their comments confirm, he believes, as evidence that he has the potential to be a competent assessment designer.

Despite the fact that his writing had improved, he discovered that, in comparison to other phases like data collection and analysis, writing was the most difficult task. He constantly went back and forth in his writing to make sure his supervisors could read it and find it to be critical, clear and concise. Nevertheless, it was not uncommon for his writing to be difficult to grasp. It normally took him several weeks or months to reflect and proceed.

However, even if it was difficult, writing also gave him opportunities to grow. He believes that *practice makes* (he prefers to say '*brings close to*') *perfect*. Accordingly, he frequently took part in UC's writing programs as well as organised the weekend writing program for the Indonesian PhD community for years. By joining this writing group, he was able to receive advice and recommendations from other top-notch members to increase his writing productivity. Thanks to this group, he eventually crossed the finish line.

The researcher's thesis submission came with an assignment to consider the meanings of the study's findings for his professional growth. In context, this study revealed the Year 9 and 12 students' SL levels, challenges and understandings. Parts of this study were presented in the webinars and published in the Mathematics Education Research Journal (MERJ); in fact, there are still materials to share and publish from the thesis. Using the findings of this study as a starting point, it is his duty to promote SL to Indonesian students. In particular, he needs to collaborate more productively with mathematics teachers to improve statistics teaching, promote the SL assessment framework to educational stakeholders, and write about SL for both domestic and international audiences. These tasks are expected to be easier given that he was appointed as one of the country coordinators of Indonesia for the International Statistical Literacy Project (ISLP), in addition to his professional background as a mathematics lecturer and researcher.

References

Abadi, & Chairani, Z. (2020). Indonesia: The development of mathematics teacher education in Indonesia. In B. R., Vogeli & M. E. A., El Tom (Eds), *Mathematics and its teaching in the Muslim world* (pp. 75-95). https://doi.org/10.1142/9789813146785_0004

Ali, N. & Peebles, D. (2013). The effect of gestalt laws of perceptual organization on the comprehension of three-variable bar and line graphs. *Human Factors*, *55*(1), 183–203. https://doi.org/10.1177%2F0018720812452592

Anagnostopoulou, K., Hatzinikita, V. & Christidou, V. (2010). Assessed students' competencies in the Greek school framework and the PISA survey. *Review of Science, Mathematics and ICT Education*, *4*(2), 43–61. https://doi.org/10.26220/rev.138

Aoyama, K. (2007). Investigating a hierarchy of students' interpretations of graphs. *International Electronic Journal of Mathematics Education*, *2*(3), 298–318. https://doi.org/10.29333/iejme/214

Aoyama, K. & Stephens, M. (2003). Graph interpretation aspects of statistical literacy: A Japanese perspective. *Mathematics Education Research Journal*, *15*(3), 207–225. https://doi.org/10.1007/BF03217380

Arteaga, P., Batanero, C., Contreras, J. M. & Cañadas, G. R. (2012). Understanding statistical graphs: a research survey. *Boletín de Estadística e Investigación Operativa*, *28*(3), 261-277.

Arteaga, P., Diaz-Levicoy, D. & Batanero, C. (2021). Primary school students' reading levels of line graphs. *Statistics Education Research Journal*, *20*(2), Article 6. https://doi.org/10.52041/serj.v20i2.339

As'ari, A. R., Chandra, T. D., Yuwono, I., Anwar, L., Nasution, S. H., Hasanah, D., Muksar, M., Sari, V. K. & Atikah, N. (2018). *Matematika—studi dan pengajaran untuk kelas 12* [Mathematics— teaching and learning for year 12]. Pusat Kurikulum dan Perbukuan, Balitbang, Kemdikbud.

As'ari, A. R., Tohir, M., Valentino, E., Imron, Z. & Taufiq, I. (2016). *Matematika—studi dan pengajaran untuk kelas 7* [Mathematics—teaching and learning for year 7]. Pusat Kurikulum dan Perbukuan, Balitbang, Kemdikbud.

As'ari, A. R., Tohir, M., Valentino, E., Imron, Z. & Taufiq, I. (2017a). *Matematika—studi dan pengajaran untuk kelas 8* [Mathematics—teaching and learning for year 8]. Pusat Kurikulum dan Perbukuan, Balitbang, Kemdikbud.

As'ari, A. R., Tohir, M., Valentino, E., Imron, Z. & Taufiq, I. (2017b). *Buku guru matematika—studi dan pengajaran untuk kelas 8* [Teacher book for mathematics—teaching and learning for year 8]. Pusat Kurikulum dan Perbukuan, Balitbang, Kemdikbud.

Azzizah, Y. (2015). Socio-economic factors on Indonesia education disparity. *International Education Studies*, *8*(12), 218–229.

Badan Pusat Statistik. (2022). *Statistik Pendidikan 2021*. https://www.bps.go.id/publication/2021/11/26/d077e67ada9a93c99131bcde/statistik-pendidikan-2021.html

Badan Pusat Statistik. (2023, 2 February). *Indonesian population*. https://www.bps.go.id/pressrelease/2021/01/21/1854/hasil-sensus-penduduk-2020.html

Badan Litbangkes, R. I (Penelitian dan Pengembangan Kesehatan Kementerian Kesehatan RI). (2013). Penyajian Pokok-Pokok Hasil Riset Kesehatan Dasar 2013. Jakarta. pp 92.

Bailey, N. G. & McCulloch, A. W. (2023). Describing critical statistical literacy habits of mind. *The Journal of Mathematical Behavior*, *70*, 101063. https://doi.org/10.1016/j.jmathb.2023.101063

Bappenas. (2015). *Data dasar pendidikan* [Basic data of education].

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R. & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematical Education in Science and Technology*, *25*(4), 527–547. https://doi.org/10.1080/0020739940250406

Beatty, P. C. & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287–311. https://doi.org/10.1093/poq/nfm006

Bedi, A. S. & Garg, A. (2000). The effectiveness of private versus public schools: The case of Indonesia. *Journal of Development Economics*, *61*(2), 463–494. https://doi.org/10.1016/S0304-3878(00)00065-1

Ben-Zvi, D. & Garfield, J. B. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi, & J. B. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3–16). Kluwer Academic Publishers.

Berndt, M., Schmidt, F. M., Sailer, M., Fischer, F., Fischer, M. R. & Zottmann, J. M. (2021). Investigating statistical literacy and scientific reasoning & argumentation in medical-, social sciences-, and economics students. *Learning and Individual Differences*, *86*, 101963. https://doi.org/10.1016/j.lindif.2020.101963

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and Quantity*, *36*(4), 391–409. https://doi.org/10.1023/A:1020909529486

Boone, H. N. & Boone, D. A. (2012). Analyzing Likert data. *Journal of Extension*, *50*(2), 1–5.

Bourque, Linda B. (2003). Cross-sectional design. In Lewis-Beck, M. S., Bryman A., & Futing Liao, T. (Eds.), *The Sage encyclopedia of social science research methods* (pp. 229). Sage Publications. https://doi.org/10.4135/9781412950589

Brown, N. J., Nagashima, S. O., Fu, A., Timms, M. & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment*, *15*(3–4), 142–174. https://doi.org/10.1080/10627197.2010.530562

Budgett, S. & Pfannkuch, M. (2010). Assessing students' statistical literacy. In P. Bidgood, N. Hunt & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 103–121). https://doi.org/10.1002/9780470710470.ch9

Budgett, S. & Renelle, A. (2023). Statistical literacy–the golden rules a review of Tim Harford's The Data Detective: Ten easy rules to make sense of statistics. *The Mathematics Enthusiast*, *20*(1), 256-265. https://doi.org/10.54870/1551-3440.1612

Budgett, S. & Rose, D. (2017). Developing statistical literacy in the final school year. *Statistics Education Research Journal*, *16*(1), 139–162. https://doi.org/10.52041/serj.v16i1.221

Bursal, M. & Yetiş, S. (2020). Middle school students' graph skills and affective states about graphs. *International Journal of Research in Education and Science (IJRES)*, *6*(4), 692-704. http://dx.doi.org/10.46328/ijres.v6i4.1136

Büscher, C. (2022a). Design principles for developing statistical literacy in middle schools. *Statistics Education Research Journal*, *21*(1), 1-16. https://doi.org/10.52041/serj.v21i1.80

Büscher, C. (2022b, February). Learning opportunities for statistical literacy in German middle school mathematics textbooks. In *Twelfth Congress of the European Society for Research in Mathematics Education*.

Callingham, R., Carmichael, C. & Watson, J. (2016). Explaining student achievement: The influence of teachers' pedagogical content knowledge in statistics. *International Journal of Science and Mathematics Education*, *14*(7), 1339–1357. https://doi.org/10.1007/s10763-015-9653-2

Callingham, R. & Watson, J. (2017). The development of statistical literacy at school. *Statistics Education Research Journal*, *16*(1), 181–201. https://doi.org/10.52041/serj.v16i1.223

Carel, G., & Juandi, D. (2023). Student statistical literacy in Indonesia: Systematic literature review. In *Proceedings of International Conference on Education of Suryakancana*. IConnects.

Carmichael, C. S. & Hay, I. (2009). Gender differences in middle school students' interests in a statistical literacy context. In *Proceedings of the 32nd Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 89–96). Palmerston North. https://researchoutput.csu.edu.au/files/9701598/Carmichael2_RP09[1].pdf

Case, C. & Jacobbe, T. (2018). A framework to characterize student difficulties in learning inference from a simulation-based approach. *Statistics Education Research Journal*, *17*(2), 9-29. https://doi.org/10.52041/serj.v17i2.156

Çatman Aksoy, E. & Işıksal Bostan, M. (2021). Seventh graders' statistical literacy: An investigation on bar and line graphs. *International Journal of Science and Mathematics Education*, *19*, 397-418. https://doi.org/10.1007/s10763-020-10052-2

Cecere, G., Corrocher, N. & Guerzoni, M. (2018). Price or performance? A probabilistic choice analysis of the intention to buy electric vehicles in European countries. *Energy Policy*, *118*, 19–32. https://doi.org/10.1016/j.enpol.2018.03.034

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3).

Chiesi, F. & Primi, C. (2015). Gender differences in attitudes toward statistics: Is there a case for a confidence gap? In *CERME 9—Ninth Congress of the European Society for Research in Mathematics Education* (pp. 622–628). https://doi.org/10.1080/10691898.2002.11910677

Choi, K. M. & Park, H. J. (2013). A comparative analysis of geometry education on curriculum standards, textbook structure, and textbook items between the US and Korea. *Eurasia Journal of Mathematics, Science and Technology Education*, *9*(4), 379–391.

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

Conrad, F. & Blair, J. (1996). From impressions to data: Increasing the objectivity of cognitive interviews. In *Proceedings of the Section on Survey Research Methods, Annual Meetings of the American Statistical Association* (Vol. 1, No. 10). American Statistical Association. https://www.bls.gov/osmr/research-papers/1996/pdf/st960080.pdf

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. Sage Publications.

Cui, Y., Chen, F., Lutsyk, A., Leighton, J. P. & Cutumisu, M. (2023). Data literacy assessments: a systematic literature review. *Assessment in Education: Principles, Policy & Practice*, *30*(1), 76-96. https://doi.org/10.1080/0969594X.2023.2182737

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education*, *18*(5), 382–393. https://doi.org/10.5951/jresematheduc.18.5.0382

da Silva, A. S., Barbosa, M. T. S., de Souza Velasque, L., da Silveira Barroso Alves, D. & Magalhães, M. N. (2021). The COVID-19 epidemic in Brazil: How statistics education may contribute to unravel the reality behind the charts. *Educational Studies in Mathematics*, *108*(1), 269–289. https://doi.org/10.1007/s10649-021-10112-6

Dahlstrom-Hakki, I. & Wallace, M. L. (2022). Teaching statistics to struggling students: Lessons learned from students with LD, ADHD, and Autism. *Journal of Statistics and Data Science Education*, *30*(2), 127-137. https://doi.org/10.1080/26939169.2022.2082601

Delport, D. H. (2023). The development of statistical literacy among students: Analyzing messages in media articles with Gal's worry questions. *Teaching Statistics*, *45*(2), 61-68. https://doi.org/10.1111/test.12308

Desimone, L. M. & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, *26*(1), 1–22. https://doi.org/10.3102%2F01623737026001001

DeWalt, D. A., Rothrock, N., Yount, S. & Stone, A. A. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care*, *45*(5 Suppl. 1), S12. https://doi.org/10.1097%2F01.mlr.0000254567.79743.e2

Dierdorp, A., Bakker, A., Ben-Zvi, D. & Makar, K. (2017). Secondary students' consideration of variability in measurement activities based on authentic practices. *Statistics Education Research Journal*, *16*(2), 397–418. https://doi.org/10.52041/serj.v16i2.198

Dwityas, N. A. & Briandana, R. (2017). Social media in travel decision making process. *International Journal of Humanities and Social Science*, *7*(7), 193–201.

Ekanayake, S., Ahmad, F. & McKenzie, K. (2012). Qualitative cross-sectional study of the perceived causes of depression in South Asian origin women in Toronto. *BMJ Open*, *2*(1), e000641. http://dx.doi.org/10.1136/bmjopen-2011-000641

Engledowl, C. & Weiland, T. (2021). Data (Mis) representation and COVID-19: Leveraging misleading data visualizations for developing statistical literacy across grades 6–16. *Journal of Statistics and Data Science Education*, *29*(2), 160-164. https://doi.org/10.1080/26939169.2021.1915215

Fakhmi, L., Sampoerno, P. & Meiliasari, M. (2021). Design research: Lintasan pembelajaran statistika untuk menumbuhkan kemampuan literasi statistik siswa. Histogram: *Jurnal Pendidikan Matematika*, *5*(2), 249-265. https://journal.stkip-andi-matappa.ac.id/index.php/histogram/article/view/1164

Ferligoj, A. (2015). How to improve statistical literacy? *Advances in Methodology and Statistics*, *12*(1), 1-10. https://doi.org/10.51936/xcxb9472

Fitriyah, L. (2020). Pengaruh PISA (Program for International Student Assessment) terhadap pendidikan di Indonesia. *Academia*.

Franklin, C. (2021). As Covid makes clear, statistics education is a must. *Significance (Oxford, England)*, *18*(2), 35. https://doi.org/10.1111/1740-9713.01509

Franklin, C. & Bargagliotti, A. (2020). Introducing GAISE II: A guideline for precollege statistics and data science education. *Harvard Data Science Review*, *2*(4), 1-9. https://doi.org/10.1162/99608f92.246107bb

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M. & Scheaffer, R. (2005). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) report*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GAISEPreK-12_Full.pdf

Freimuth, H. (2016). An examination of cultural bias in IELTS Task 1 non-process writing prompts: A UAE perspective. *Learning and Teaching in Higher Education: Gulf Perspectives*, *13*(1). https://www.emerald.com/insight/content/doi/10.18538/lthe.v13.n1.221/full/html

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*(1), 1–25. https://doi.org/10.2307/1403713

Gal, I. (2019). Understanding statistical literacy: About knowledge of contexts and models. In J. M. Contreras, M. M. Gea, M. M. López-Martín & E. Molina-Portillo (Eds.), *Actas del Tercer Congreso Internacional Virtual de Educación Estadística*. https://www.ugr.es/local/fqm126/civeest.html

Gal, I. & Geiger, V. (2022). Welcome to the era of vague news: A study of the demands of statistical and mathematical products in the COVID-19 pandemic media. *Educational Studies in Mathematics*, *111*(1), 5–28. https://doi.org/10.1007/s10649-022-10151-7

Garfield, J., del Mas, R. & Zieffler, A. (2010). Assessing important learning outcomes in introductory tertiary statistics courses. In P. Bidgood, N. Hunt & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 75–86). https://doi.org/10.1002/9780470710470.ch7

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

Glaser, B. G. & Strauss, A. L. (2017). *The discovery of grounded theory: Strategies for qualitative research*. Routledge. https://doi.org/10.4324/9780203793206

Glasnovic Gracin, D. (2018). Requirements in mathematics textbooks: A five-dimensional analysis of textbook exercises and examples. *International Journal of Mathematical Education in Science and Technology*, *49*(7), 1003–1024. https://doi.org/10.1080/0020739X.2018.1431849

Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, *47*(2), 183–210. https://doi.org/10.1080/03057267.2011.605307

Gonda, D., Pavlovičová, G., Ďuriš, V. & Tirpáková, A. (2022). Implementation of pedagogical research into statistical courses to develop students' statistical literacy. *Mathematics*, *10*(11), 1793. https://doi.org/10.3390/math10111793

Groth, R. E. (2014). Using work samples from the National Assessment of Educational Progress (NAEP) to design tasks that assess statistical knowledge for teaching. *Journal of Statistics Education*, *22*(3). https://doi.org/10.1080/10691898.2014.11889712

Groth, R. E. & Bergner, J. A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning*, *8*(1), 37-63. https://doi.org/10.1207/s15327833mtl0801_3

Guler, M., Gursoy, K. & Guven, B. (2016). Critical views of 8th Grade students toward statistical data in newspaper articles: Analysis in light of statistical literacy. *Cogent Education*, *3*(1), 1268773. https://doi.org/10.1080/0020739X.2018.1431849

Hafiyusholeh, M. (2015). Literasi statistik dan urgensinya bagi siswa. *Wahana*, *64*(1), 1-8.

Hafiyusholeh, M., Budayasa, K. & Siswono, T. (2018). Statistical literacy: High school students in reading, interpreting and presenting data. *Journal of Physics: Conference Series*, *947*. https://doi.org/10.1088/1742-6596/947/1/012036

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, *7*(6), 836–850. https://doi.org/10.1016/j.cptl.2015.08.001

Hasan, T., Muhaddes, T., Camellia, S., Selim, N. & Rashid, S. F. (2014). Prevalence and experiences of intimate partner violence against women with disabilities in Bangladesh: Results of an explanatory sequential mixed-method study. *Journal of Interpersonal Violence*, *29*(17), 3105–3126. https://doi.org/10.1177/0886260514534525

Helenius, R., D’Amelio, A., Campos, P. & Macfeely, S. (2020). ISLP country coordinators as ambassadors of statistical literacy and innovations. *Statistics Education Research Journal*, *19*(1), 120-136. https://doi.org/10.52041/serj.v19i1.125

Hoffrén, J. (2021). Statistical literacy competencies as post-modern basic civic skills. In *Proceedings 63rd ISI World Statistics Congress* (Vol. 11, p. 16).

Inspektorat Jenderal Kemdikbud. (2023, 2 February). *Curriculum changes in Indonesia*. https://itjen.kemdikbud.go.id/

Irwandi, B., Roza, Y. & Maimunah, M. (2022). Analisis kemampuan literasi statistis peserta Asesmen Kompetensi Minimum (AKM). *Jurnal Gantang*, *6*(2), 177–183. https://doi.org/10.31629/jg.v6i2.3961

Jacobbe, T. & Carvalho, C. (2011). Teachers’ understanding of averages. In C. Batanero, G. Burrill, C. Reading, & A. Rossman (Eds.), *Teaching statistics in school mathematics challenges for teaching and teacher education* (pp. 199–209). Springer.

Johannssen, A., Chukhrova, N., Schmal F. & Stabenow, Kevin. (2021). Statistical literacy—misuse of statistics and its consequences. *Journal of Statistics and Data Science Education*, *29*(1), 54-62. https://doi.org/10.1080/10691898.2020.1860727

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B. & Putt, I. J. (2000). A framework for characterizing children’s statistical thinking. *Mathematical Thinking and Learning*, *2*(4), 269–307. https://doi.org/10.1207/S15327833MTL0204_3

Joram, E., Resnick, L. B. & Gabriele, A. J. (1995). Numeracy as cultural practice: An examination of numbers in magazines for children, teenagers, and adults. *Journal for Research in Mathematics Education*, *26*(4), 346–361. https://doi.org/10.5951/jresematheduc.26.4.0346

Joshi, A., Kale, S., Chandel, S. & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, *7*(4), 396. https://doi.org/10.9734/BJAST/2015/14975

Jupri, A. (2015). *The use of applets to improve Indonesian student performance in algebra* [Unpublished doctoral dissertation]. Utrecht University. https://dspace.library.uu.nl/handle/1874/303950

Jurdak, M. E., Mouhayar, E. L. & Raif, R. (2014). Trends in the development of student level of reasoning in pattern generalization tasks across grade level. *Educational Studies in Mathematics*, *85*(1), 75–92. https://doi.org/10.1007/s10649-013-9494-2

Jureckova, M. & Csachova, L. (2020). Statistical literacy of Slovak lower secondary school students. *Technium Social Sciences Journal*, *9*, 163-173. http://dx.doi.org/10.47577/tssj.v9i1.966

Kemdikbud. (2012). *Dokumen Kurikulum 2013* [Document of Curriculum 2013].

Kemdikbud. (2013). *Kurikulum 2013. Kompetensi dasar: Sekolah Menengah Pertama (SMP)/Madrasah Tsanawiyah (MTs)* [Curriculum 2013. Basic Competencies: Junior Secondary School].

Kemdikbud. (2016). *Indonesia educational statistics in brief 2015-2016*. Centre for educational data and statistics and culture.

Kemdikbud. (2023, 2 February). *The number of school students*. https://statistik.data.kemdikbud.go.id/

Kemp, M. & Kissane, B. (2010, 11–16 July). *A five step framework for interpreting tables and graphs in their contexts* [Paper presentation]. 8th International Conference on Teaching Statistics, Ljubljana, Slovenia. https://researchrepository.murdoch.edu.au/id/eprint/6240/

Kesmodel, U. S. (2018). Cross-sectional studies—what are they good for? *Acta Obstetricia et Gynecologica Scandinavica*, *97*(4), 388–393. https://doi.org/10.1111/aogs.13331

Kharismawati, S. A. (2022). Evaluasi Pelaksanaan Asesmen Nasional Berbasis Komputer di Sekolah Dasar Terpencil [Evaluation of the implementation of the computer-based National Assessment in remote elementary schools]. *Ideguru: Jurnal Karya Ilmiah Guru*, *7*(2), 229–234.

Klein, T., Galdin, A. & Mohamedou, E. (2016, July). An indicator for statistical literacy based on national newspaper archives. In *Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE)*.

Koga, S. (2022a). Characteristics of statistical literacy skills from the perspective of critical thinking. *Teaching Statistics*, *44*(2), 59-67. https://doi.org/10.1111/test.12302

Koga, S. (2022b). Lessons aimed at demonstrating statistical literacy skills: A case study of Japanese high school lessons on reading statistical reports. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics.* International Association for Statistical Education. https://doi.org/10.52041/iase.icots11.T7B1

Koparan, T. & Güven, B. (2015). The effect of project-based learning on students' statistical literacy levels for data representation. *International Journal of Mathematical Education in Science and Technology*, *46*(5), 658–686. https://doi.org/10.1080/0020739X.2014.995242

Krishnan, S. (2015). Fostering students' statistical literacy through significant learning experience. *Journal of Research in Mathematics Education*, *4*(3), 259–270. https://doi.org/10.17583/redimat.2015.1332

Kurnia, A. B., Lowrie, T. & Patahuddin, S. M. (2023). The development of high school students' statistical literacy across grade level. *Mathematics Education Research Journal*, 1-29. https://doi.org/10.1007/s13394-023-00449-x

Laerd Statistics. (2015). Mann-Whitney *U* test using SPSS Statistics. *Statistical Tutorials and Software Guides*. https://statistics.laerd.com/

Laerd Statistics. (2016). Kendall's coefficient of concordance, *W*, using SPSS Statistics. *Statistical Tutorials and Software Guides*. https://statistics.laerd.com/

Landtblom, K. (2018). Is data a quantitative thing? An analysis of the concept of the mode in textbooks for grade 4-6. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward*. International Statistical Institute. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_2F1.pdf

Langrall, C., Nisbet, S., Mooney, E. & Jansem, S. (2011). The role of context expertise when comparing data. *Mathematical Thinking and Learning*, *13*(1–2), 47–67. https://doi.org/10.1080/10986065.2011.538620

Leavy, A. & O'Loughlin, N. (2006). Preservice teachers' understanding of the mean: Moving beyond the arithmetic average. *Journal of Mathematics Teacher Education*, *9*(1), 53–90. https://doi.org/10.1007/s10857-006-9003-y.

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

Levin, K. A. (2006). Study design III: Cross-sectional studies. *Evidence-based Dentistry*, *7*(1), 24–25. https://doi.org/10.1038/sj.ebd.6400375

Li, L., Wang, Z., Chen, L. & Wang, Z. (2020). Consumer preferences for battery electric vehicles: A choice experimental survey in China. *Transportation Research Part D: Transport and Environment*, *78*, 102185. https://doi.org/10.1016/j.trd.2019.11.014

Lowrie, T. & Diezmann, C. M. (2009). National numeracy tests: A graphic tells a thousand words. *Australian Journal of Education*, *53*(2), 141–158. https://doi.org/10.1177%2F000494410905300204

Ludewig (2018). *Understanding graphs: Modeling processes, prerequisites and influencing factors of graphicacy* [Doctoral Dissertation]. Tübingen University. https://tobias-lib.ub.uni-tuebingen.de/xmlui/handle/10900/84624

Ludewig, U., Lambert, K., Dackermann, T., Scheiter, K. & Möller, K. (2020). Influences of basic numerical abilities on graph reading performance. *Psychological Research, 84*, 1198-1210. https://doi.org/10.1007/s00426-019-01144-y

Ludwig, M. & Xu, B. (2010). A comparative study of modelling competencies among Chinese and German students. *Journal für Mathematik-Didaktik*, *31*(1), 77–97. https://doi.org/10.1007/s13138-010-0005-z

Magaldi, D., & Berler, M. (2020). Semi-structured interviews. In: Zeigler-Hill, V., Shackelford, T.K. (Eds), *Encyclopedia of personality and individual differences* (pp. 4825–4830). Springer, Cham. https://doi.org/10.1007/978-3-319-24612-3_857

Mann, C. J. (2003). Observational research methods. Research design II: Cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, *20*(1), 54–60.

Marchy, F. & Juandi, D. (2023). Student's statistical literacy skills (1980-2023): A systematic literature review with bibliometric analysis. *Journal of Education and Learning Mathematics Research (JELMaR)*, *4*(1), 31-45. https://doi.org/10.37303/jelmar.v4i1.105

Merriman, L. (2006). *Using media reports to develop statistical literacy in Year 10 students* [Paper presentation]. 7th International Conference on Teaching Statistics, Salvador, Brazil.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

Montague, M. & Van Garderen, D. (2003). A cross-sectional study of mathematics achievement, estimation skills, and academic self-perception in students of varying ability. *Journal of Learning Disabilities*, *36*(5), 437–448. https://doi.org/10.1177%2F00222194030360050501

Mooney, E. S. (2002). A framework for characterizing middle school students' statistical thinking. *Mathematical Thinking and Learning*, *4*(1), 23–63. https://doi.org/10.1207/S15327833MTL0401_2

Moritz, J. (2003). Constructing coordinate graphs: Representing corresponding ordered values with variation in two-dimensional space. *Mathematics Education Research Journal*, *15*(3), 226–251.

Mullis, I. V. & Martin, M. O. (2017). *TIMSS 2019 assessment frameworks*. International Association for the Evaluation of Educational Achievement.

Mullis, I. V., Martin, M. O., Foy, P. & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* International Association for the Evaluation of Educational Achievement.

Mullis, I. V., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y. & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. International Association for the Evaluation of Educational Achievement.

Mulya, N. (2018). *Kemampuan Literasi Statistik dan Resiliensi Siswa SMP dalam Menyelesaikan Soal PISA* (Doctoral dissertation, Universitas Pendidikan Indonesia).

Muñiz-Rodríguez, L., Rodríguez-Muñiz, L. J., & Alsina, Á. (2020). Deficits in the statistical and probabilistic literacy of citizens: Effects in a world in crisis. *Mathematics*, *8*(11), 1872. https://doi.org/10.3390/math8111872

Muttaqin, H., Putri, R. I. I. & Somakim. (2017). Design research on ratio and proportion learning by using ratio table and graph with OKU Timur context at the 7th Grade. *Journal on Mathematics Education*, *8*(2), 211–222.

Muttaqin, T., Wittek, R., Heyse, L. & van Duijn, M. (2020). The achievement gap in Indonesia? Organizational and ideological differences between private Islamic schools. *School Effectiveness and School Improvement*, *31*(2), 212–242. https://doi.org/10.1080/09243453.2019.1644352

Newhouse, D. & Beegle, K. (2006). The effect of school type on academic achievement evidence from Indonesia. *Journal of Human Resources*, *41*(3), 529–557. https://doi.org/10.3368/jhr.XLI.3.529

Noor, I. H., Sabon, S. S., Joko, B. S. & Wijayanti, K. (2020a). Pengelolaan musyawarah guru mata pelajaran (MGMP) untuk memperkuat kompetensi guru. Kemdikbud. http://puslitjakdikbud.kemdikbud.go.id/

Noor, I. H., Sabon, S. S., Joko, B. S., & Wijayanti, K. (2020b). Peran musyawarah guru mata pelajaran (MGMP) dalam meningkatkan mutu pembelajaran di SMA. Kemdikbud. http://puslitjakdikbud.kemdikbud.go.id/

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. OECD Publishing.

OECD. (2013a). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing.

OECD. (2013b). *PISA 2012 released mathematics items*. https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf

OECD. (2014, February). *PISA 2012 results: What students know and can do—student performance in mathematics, reading and science* (Vol. I, rev. ed.). OECD Publishing. https://doi.org/10.1787/9789264208780-en

OECD. (2018). *PISA 2021 mathematics framework* (second draft). OECD Publishing.

OECD. (2023). *PISA 2022 results (volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en.

Oktiviani, F. N. (2021). Kemampuan literasi statistik dalam pembelajaran berbasis proyek yang dimodifikasi. *Indonesian Journal of Islamic Studies*, *2*(2), 235-247. http://journal.civiliza.org/index.php/ijois/article/view/44

Olande, O. (2014). Graphical artefacts: Taxonomy of students' response to test items. *Educational Studies in Mathematics*, *85*(1), 53–74. https://doi.org/10.1007/S10649-013-9493-3

Onwuegbuzie, A. J. & Collins, K. M. (2007). A typology of mixed methods sampling designs in social science research. *The Qualitative Report*, *12*(2), 281–316.

Padmi, R. S., Hidayati, F. H. & Hasanah, M. Indonesian students'experience with real-life data and cross-border collaboration: what does the graph say?. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward*. International Statistical Institute.

Pallauta, J. D., Arteaga, P. & Garzón-Guerrero, J. A. (2021). Secondary school students' construction and interpretation of statistical tables. *Mathematics*, *9*(24), 3197. https://doi.org/10.3390/math9243197

Parmar, R. S. & Signer, B. R. (2005). Sources of error in constructing and interpreting graphs: A study of fourth-and fifth-grade students with LD. *Journal of Learning Disabilities*, *38*(3), 250–261.

Patahuddin, S. M. & Lowrie, T. (2019). Examining teachers' knowledge of line graph task: A case of travel task. *International Journal of Science and Mathematics Education*, *17*(4), 781–800. https://doi.org/10.1007/s10763-018-9893-z

Patahuddin, S. M., Suwarsono, S. & Johar, R. (2018). Indonesia: History and perspective on mathematics education. In J. M. Mack & B. R. Vogeli (Eds), *Mathematics and its teaching in the Asia-Pacific region* (pp. 191–230). World Scientific Publishing.

Peebles, D. & Ali, N. (2015). Expert interpretation of bar and line graphs: The role of graphicacy in reducing the effect of graph format. *Frontiers in Psychology*, *6*, 1673. https://doi.org/10.3389/fpsyg.2015.01673

Perez, L. R., Spangler, D. A. & Franklin, C. (2021). Engaging young learners with data: Highlights from GAISE II, Level A. *Harvard Data Science Review*, *3*(2). https://doi.org/10.1162/99608f92.be3c2ec8

Pfannkuch, M. (2005). Characterizing Year 11 students' evaluation of a statistical process. *Statistics Education Research Journal*, *4*(2), 5–25. https://doi.org/10.52041/serj.v4i2.512

Ponte, J. P. D. & Marques, S. (2011). Proportion in school mathematics textbooks: A comparative study. *RIPEM—International Journal for Research in Mathematics Education*, *1*(1), 36–53. http://hdl.handle.net/10451/4222

Power, D. J. & Phillips-Wren, G. (2011). Impact of social media and Web 2.0 on decision-making. *Journal of Decision Systems*, *20*(3), 249–261. https://doi.org/10.3166/jds.20.249-261

Pratiwi, I. (2019). Efek program PISA terhadap kurikulum di Indonesia [The effect of the PISA program on the curriculum in Indonesia]. *Jurnal pendidikan dan Kebudayaan*, *4*(1), 51–71. https://doi.org/10.24832/jpnk.v4i1.1157

Priyambodo, S. & Maryati, I. (2019). Peningkatan kemampuan literasi statistis melalui model pembelajaran berbasis proyek yang dimodifikasi. *Mosharafa: Jurnal Pendidikan Matematika*, *8*(2), 273-284. https://doi.org/10.31980/mosharafa.v8i2.496

Pusat Penilaian Pendidikan Kemdikbud. (2023, 2 February). *The results of UN*. https://hasilun.pusmenjar.kemdikbud.go.id/

Ralston, N. C., Li, M. & Taylor, C. (2018). The development and initial validation of an assessment of algebraic thinking for students in the elementary grades. *Educational Assessment*, *23*(3), 211–227. https://doi.org/10.1080/10627197.2018.1483191

Raynes-Greenow, C. H., Gordon, A., Li, Q. & Hyett, J. A. (2013). A cross-sectional study of maternal perception of fetal movements and antenatal advice in a general pregnant population, using a qualitative framework. *BMC Pregnancy and Childbirth*, *13*(1), 1-8. https://doi.org/10.1186/1471-2393-13-32

Reinhart, A., Evans, C., Luby, A., Orellana, J., Meyer, M., Wieczorek, J., Elliot, P., Burckhardt, P. & Nugent, R. (2022). Think-aloud interviews: A tool for exploring student statistical reasoning. *Journal of Statistics and Data Science Education*, *30*(2), 100-113. https://doi.org/10.1080/26939169.2022.2063209

Reys, B. J., Reys, R. E. & Chavez, O. (2004). Why mathematics textbooks matter. *Educational Leadership*, *61*(5), 61–66.

Robinson, O. C. (2014). Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative Research in Psychology*, *11*(1), 25–41. https://doi.org/10.1080/14780887.2013.801543

Rotaru, M. C. (2018). The importance of visual literacy: An analysis of potential obstacles for Romanian students in the completion of IELTS Academic Writing Task 1. In L.-M. Grosu-Rădulescu (Ed.), *Foreign language teaching in Romanian higher education* (pp. 61–82). Springer.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3). https://doi.org/10.1080/10691898.2002.11910678

Sabbag, A., Garfield, J. & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning: The REALI instrument. *Statistics Education Research Journal*, *17*(2), 141-160. https://doi.org/10.52041/serj.v17i2.163

Sari, Y. M. & Valentino, E. (2017). An analysis of students error in solving PISA 2012 and its scaffolding. *JRAMathEdu (Journal of Research and Advances in Mathematics Education)*, *1*(2), 90–98.

Schield, M. (2000). Statistical literacy: Difficulties in describing and comparing rates and percentages. In *ASA Proceedings of the Section on Statistical Education* (p. 176).

Schield, M. (2017). GAISE 2016 promotes statistical literacy. *Statistics Education Research Journal*, *16*(1), 50-54. http://iase-web.org/Publications.php?p=SERJ

Setia, M. S. (2016). Methodology series module 3: Cross-sectional studies. *Indian Journal of Dermatology*, *61*(3), 261. https://doi.org/10.4103%2F0019-5154.182410

Shafer, K. & Lohse, B. (2005). *How to conduct a cognitive interview: A nutrition education example*. National Institute of Food and Agriculture, U.S. Department of Agriculture.

Sharma, S. (2013a). Assessıng students' understandıng of tables and graphs: implıcatıons for teachıng and research. *International Journal of Educational Research and Technology*, *4*(4), 61-69.

Sharma, S. (2013b). Developing statistical literacy with Year 9 students: a collaborative research project. *Research in Mathematics Education*, *15*(2), 203-204. https://doi.org/10.1080/14794802.2013.797742

Sharma, S. (2014). Influence of culture on secondary school students' understanding of statistics: A Fijian perspective. *Statistics Education Research Journal*, *13*(2), 104–117. https://doi.org/10.52041/serj.v13i2.284

Sharma, S (2017) Definitions and models of statistical literacy: a literature review. *Open Review of Educational Research*, 4:1, 118-133. https://doi.org/10.1080/23265507.2017.1354313

Sharma, S. (2018a). Bridging language barriers in statistics for Year-12 Pasifika students: A collaborative study. *Waikato Journal of Education*, *23*(1), 107–120. https://doi.org/10.15663/wje.v23i1.646

Sharma, S. (2018b). Enhancing statistical literacy through real world examples: a collaborative study. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward*. International Statistical Institute. http://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_9D3.pdf

Sharma, S. V. (2006). High school students interpreting tables and graphs: Implications for research. *International Journal of Science and Mathematics Education*, *4*(2), 241–268. https://doi.org/10.1007/s10763-005-9005-8

Sharma, S., Doyle, P., Shandil, V., & Talakia'atu, S. (2011). Developing statistical literacy with Year 9 students. *Set: Research Information for Teachers*, (1), 43-50.

Sharma, S., Doyle, P., Shandil, V. & Talakia'atu, S. (2012). A four-stage framework for assessing statistical literacy. *Curriculum Matters*, *8*, 148–170. https://doi.org/10.18296/cm.0139

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Eds.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp 957-1009). Reston: The National Council of Teachers of Mathematics.

Sherington, J., Stern, R. D., Allan, E. F., Wilson, I. M. & Coe, R. (2004). Informative presentation of tables, graphs and statistics. In R. D. Stern, R. Coe, E. F. Allan & I. C. Dale (Eds.), *Statistical good practice for natural resources research*. CAB International.

Shields, M. (2005). Information literacy, statistical literacy, data literacy. *IASSIST Quarterly, 28*(2–3), 6. https://doi.org/10.29173/iq790

Sinaga, B., Sinambela, P. N. J. M., Sitanggang, A. K., Hutapea, T. A., Sinaga, L. P., Manullang, S., Simanjorang, M. & Bayuzetra, Y. T. (2014a). *Matematika—studi dan pengajaran untuk kelas 10* [Mathematics— teaching and learning for year 10]. Pusat Kurikulum dan Perbukuan, Balitbang, Kemdikbud.

Sinaga, B., Sinambela, P. N. J. M., Sitanggang, A. K., Hutapea, T. A., Sinaga, L. P., Manullang, S., Simanjorang, M. & Bayuzetra, Y. T. (2014b). *Matematika—studi dan pengajaran untuk kelas 11* [Mathematics—teaching and learning for year 11]. Pusat Kurikulum dan Perbukuan, Balitbang, Kemdikbud.

Sinaga, B., Sinambela, P. N. J. M., Sitanggang, A. K., Hutapea, T. A., Sinaga, L. P., Manullang, S., Simanjorang, M. & Bayuzetra, Y. T. (2014c). *Buku guru matematika—studi dan pengajaran untuk kelas 10* [Teacher book for mathematics—teaching and learning for year 10]. Pusat Kurikulum dan Perbukuan, Balitbang, Kemdikbud.

Snider, V. E. (2004). A comparison of spiral versus strand curriculum. *Journal of Direct Instruction*, *4*(1), 29-39.

Spector, Paul E. (2003). Cross-sectional data. In M. Lewis-Beck, A. E. Bryman & T. F. Liao (Eds). *The Sage encyclopedia of social science research methods* (pp. 229–230). Sage Publications.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Suarez-Alvarez, J. (2021). Are 15-year-olds prepared to deal with fake news and misinformation?. *PISA in Focus*, No. 113, OECD Publishing. https://doi.org/10.1787/6ad5395e-en

Subchan, Winarni, Hanafi, L., Mufid, M. S., Fahim, K., Syaifudin, W. H. & Cahyaningtias, S. (2015). *Matematika—studi dan pengajaran untuk kelas 9* [Mathematics—teaching and learning for year 9]. Pusat Kurikulum dan Perbukuan, Balitbang, Kemdikbud.

Sumaryanta, S., Priatna, N. & Sugiman, S. (2019). Pemetaan hasil ujian nasional matematika [Mapping the results of the national mathematics exam]. *Idealmathedu: Indonesian Digital Journal of Mathematics and Education*, *6*(1), 543–557. https://doi.org/10.53717/idealmathedu.v6i1.38

Suri, H. (2011). Purposeful sampling in qualitative research synthesis. *Qualitative Research Journal*, *11*(2), 63–75. https://doi.org/10.3316/QRJ1102063

Susanto, D., Sihombing, S. K., Radjawane, M. M., Candra, Y. & Sinambela, D. (2021). *Matematics for Year 11*. Pusat Perbukuan, Badan Standar Kurikulum dan Asesmen Pendidikan. Kemendikbudristek. https://buku.kemdikbud.go.id/katalog/matematika-untuk-smasmk-kelas-xi

Sutherland, M., Fainstein, D., Lesner, T., Kimmel, G. L., Clarke, B. & Doabler, C. T. (2022). Teaching statistical literacy and data analysis to students with mathematics difficulties. *Teaching Exceptional Children*, 1-11. https://doi.org/10.1177/00400599221118647

Sutton, S. (2000). Interpreting cross-sectional data on stages of change. *Psychology and Health*, *15*(2), 163–171. https://doi.org/10.1080/08870440008400298

Taniyama, M., Kai, I. & Takahashi, M. (2012). Differences and commonalities in difficulties faced by clinical nursing educators and faculty in Japan: a qualitative cross-sectional study. *BMC Nursing*, *11*, 1-11. https://doi.org/10.1186/1472-6955-11-21

Tarran, B. (2017). How to measure statistical literacy? *Significance*, *14*(1), 42–43.

The Regulation of the Minister of National Education, Culture, Research and Technology. (2021). No.17, Year 2021 about *Asesmen Nasional* (National Assessment). jdih.kemdikbud.go.id.

Tim Gakko Tosho. (2022). *Matematics for Junior High School Year 9*. Pusat Perbukuan, Badan Standar Kurikulum dan Asesmen Pendidikan. Kemendikbudristek. https://buku.kemdikbud.go.id/katalog/matematika-untuk-sekolah-menengah-pertama-kelas-ix

TIMSS & PIRLS. (2011). *TIMSS mathematics released items*. https://timssandpirls.bc.edu/timss2011/international-released-items.html

Tiro, M. A. (2017, March). Teaching statistics in Indonesia school: Today and future. In *ISI Regional Statistics Conference* (pp. 1-10).

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J. & Janssen, R. (2017). Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. *International Journal of Science and Mathematics Education*, *15*(4), 663–682. https://doi.org/10.1007/s10763-016-9723-0

Valentina, K. (2016). Graph description as an issue in L2 academic English writing. *Journal of Language and Education*, *2*(4), 46–54.

Valverde, G. A., Bianchi, L. J., Wolfe, R. G., Schmidt, W. H. & Houang, R. T. (2002). *According to the book: Using TIMSS to investigate the translation of policy into practice through the world of textbooks*. Springer.

Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education*. Utrecht University. https://dspace.library.uu.nl/handle/1874/1705

von Roten, F. C. & de Roten, Y. (2013). Statistics in science and in society: From a state-of-the-art to a new research agenda. *Public Understanding of Science*, *22*(7), 768–784. https://doi.org/10.1177/0963662513495769

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, *88*(421), 1–8. https://doi.org/10.1080/01621459.1993.10594283

Wang, X. & Cheng, Z. (2020). Cross-sectional studies: Strengths, weaknesses, and recommendations. *Chest*, *158*(1), S65–S71. https://doi.org/10.1016/j.chest.2020.03.012

Wang, Z., & McDougall, D. (2019). Curriculum matters: What we teach and what students gain. *International Journal of Science and Mathematics Education*, *17*, 1129-1149. https://doi.org/10.1007/s10763-018-9915-x

Watson, J.M. (1997). Assessing statistical thinking using the media. In I. Gal & J.B. Garfield (Eds.), The assessment challenge in statistics education (pp. 107-121). IOS Press and The International Statistical Institute.

Watson, J. (2006). *Statistical literacy at school: Growth and goals*. Lawrence Erlbaum Association.

Watson, J. & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3–46. https://doi.org/10.52041/serj.v2i2.553

Watson, J. & Callingham, R. (2014). Two-way tables: Issues at the heart of statistics and probability for students and teachers. *Mathematical Thinking and Learning*, *16*(4), 254–284. https://doi.org/10.1080/10986065.2014.953019

Watson, J. & Callingham, R. (2020). COVID-19 and the need for statistical literacy. *Australian Mathematics Education Journal*, *2*(2), 16–21.

Watson, J. M. & Moritz, J. B. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning*, *2*(1–2), 11–50. https://doi.org/10.1207/S15327833MTL0202_2

Weathington, B. L., Cunningham, C. J. & Pittenger, D. J. (2010). *Research methods for the behavioral and social sciences*. Wiley.

Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. *Educational Studies in Mathematics*, *96*(1), 33–47. https://doi.org/10.1007/s10649-017-9764-5

Weiland, T. (2019). The contextualized situations constructed for the use of statistics by school mathematics textbooks. *Statistics Education Research Journal*, *18*(2), 18-38. https://doi.org/10.52041/serj.v18i2.138

Weiland, T. & Sundrani, A. (2022). Towards a framework for developing a critical statistical literacy. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics.* International Association for Statistical Education. https://doi.org/10.52041/iase.icots11.T1C3

West, J. D. & Bergstrom, C. T. (2020). *Calling bullshit: the art of scepticism in a data-driven world*. Penguin UK.

Whalley, L. J. (2006). Nutrients and aging. In M. Conn (Ed.), *Handbook of models of human aging* (1st ed., pp. 897–911). Elsevier.

Whitacre, I., Azuz, B., Lamb, L. L., Bishop, J. P., Schappelle, B. P. & Philipp, R. A. (2017). Integer comparisons across the grades: Students' justifications and ways of reasoning. *The Journal of Mathematical Behavior*, *45*, 47–62. https://doi.org/10.1016/j.jmathb.2016.11.001

Whitaker, D., Foti, S. & Jacobbe, T. (2015). The Levels of Conceptual Understanding in Statistics (LOCUS) Project: Results of the pilot study. *Numeracy: Advancing Education in Quantitative Literacy*, *8*(2). https://digitalcommons.usf.edu/numeracy/vol8/iss2/art3/

Wijaya, A., Retnawati, H., Setyaningrum, W. & Aoyama, K. (2019). Diagnosing students' learning difficulties in the eyes of Indonesian mathematics teachers. *Journal on Mathematics Education*, *10*(3), 357–364.

Wijaya, A., van den Heuvel-Panhuizen, M., Doorman, M. & Robitzsch, A. (2014). Difficulties in solving context-based PISA mathematics tasks: An analysis of students' errors. *The Mathematics Enthusiast*, *11*(3), 555–584.

Wild, C. J. (2017). Statistical literacy as the earth moves. *Statistics Education Research Journal*, *16*(1), 31-37. https://doi.org/10.52041/serj.v16i1.211

Willis, G. B. (1999). *Cognitive interviewing: A 'how to' guide*. Research Triangle Institute.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Sage.

Winarti, D. W. & Patahuddin, S. M. (2017). *Graphic-rich items within high-stakes tests: Indonesia National Exam (UN), PISA, and TIMSS*. Mathematics Education Research Group of Australasia.

Woodard, V., Lee, H. & Woodard, R. (2020). Writing assignments to assess statistical thinking. *Journal of Statistics Education*, *28*(1), 32-44. https://doi.org/10.1080/10691898.2019.1696257

Yilmaz, Z., Ergül, K. & Aşik, G. (2023). Role of context in statistics: Interpreting social and historical events. *Statistics Education Research Journal*, *22*(1), Article 6. https://doi.org/10.52041/serj.v22i1.72

Yolcu, A. (2014). Middle school students' statistical literacy: Role of grade level and gender. *Statistics Education Research Journal*, *13*(2), 118–131. https://doi.org/10.52041/serj.v13i2.285

Yorke, M. & Zaitseva, E. (2013). Do cross-sectional student assessment data make a reasonable proxy for longitudinal data? *Assessment & Evaluation in Higher Education*, *38*(8), 957–967. https://doi.org/10.1080/02602938.2013.769199

Yotongyos, M., Traiwichitkhun, D. & Kaemkate, W. (2015). Undergraduate students' statistical literacy: A survey study. *Procedia—Social and Behavioral Sciences, 191,* 2731–2734.

Yusuf, Y., Suyitno, H., & Sukestiyarno, Y. L. (2020, June). The identification of pre-service mathematics teachers' statistical reasoning on descriptive statistics. In *International Conference on Science and Education and Technology (ISET 2019)* (pp. 105-109). Atlantis Press. https://doi.org/10.2991/assehr.k.200620.021

Zawojewski, J. S. & Shaughnessy, J. M. (2000). Take time for action: Mean and median: Are they really so easy? *Mathematics Teaching in the Middle School*, *5*(7), 436–440.

Zheng, M. (2015). Conceptualization of cross-sectional mixed methods studies in health science: A methodological review. *International Journal of Quantitative and Qualitative Research Methods*, *3*(2), 66–87.

Zieffler, A., Garfield, J. & Fry, E. (2018). What is statistics education? In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education* (pp. 37–70). Springer. https://doi.org/10.1007/978-3-319-66195-7_2

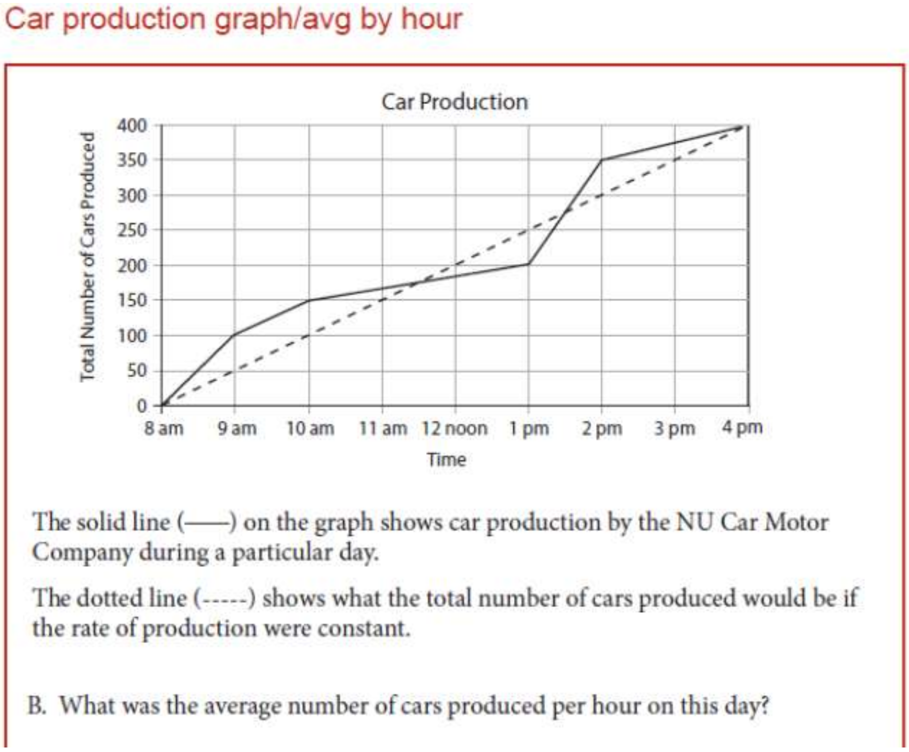
Ziegler, L. & Garfield, J. (2018). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal*, *17*(2), 161–178. https://doi.org/10.52041/serj.v17i2.164

Zuhdi, M. (2015). Pedagogical Practices in Indonesia. In E. H., Law & U. Miura (Eds), *Transforming teaching and learning in Asia and the Pacific case studies from seven countries* (pp. 142 - 160).

Zulkardi & Putri, R. I. I. (2019). New school mathematics curricula, PISA and PMRI in Indonesia. In C. Vistro-Yu & T. Toh (Eds.), *School mathematics curricula. Mathematics education—an Asian perspective*. Springer. https://doi.org/10.1007/978-981-13-6312-2

Appendix A. Original Fourteen Items

Car production graph 1



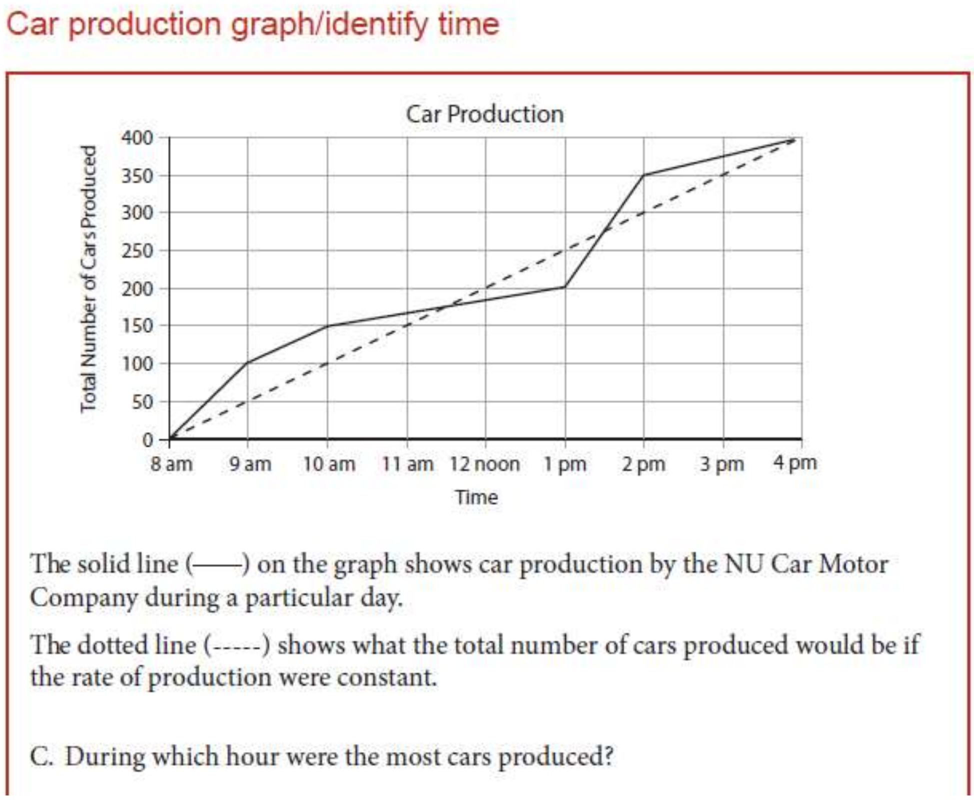
Car production graph/avg by hour

The solid line (——) on the graph shows car production by the NU Car Motor Company during a particular day.

The dotted line (-----) shows what the total number of cars produced would be if the rate of production were constant.

B. What was the average number of cars produced per hour on this day?

Car production graph 2



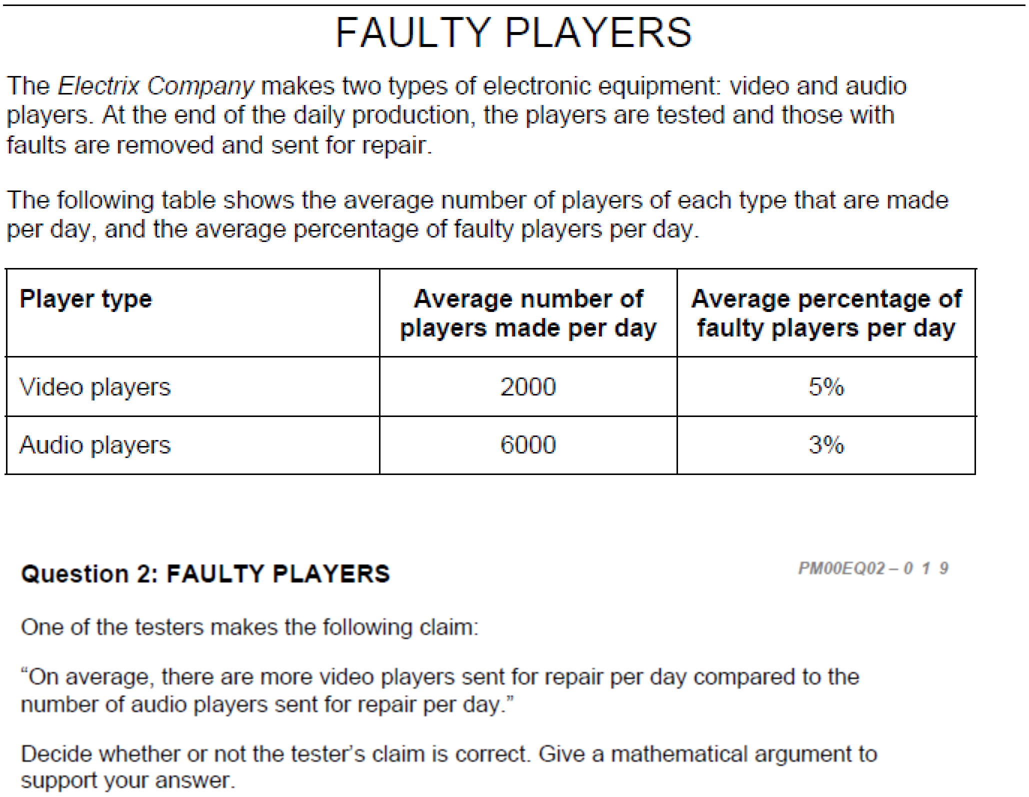
Car production graph/identify time

The solid line (——) on the graph shows car production by the NU Car Motor Company during a particular day.

The dotted line (-----) shows what the total number of cars produced would be if the rate of production were constant.

C. During which hour were the most cars produced?

Faulty players



FAULTY PLAYERS

The *Electrix Company* makes two types of electronic equipment: video and audio players. At the end of the daily production, the players are tested and those with faults are removed and sent for repair.

The following table shows the average number of players of each type that are made per day, and the average percentage of faulty players per day.

Player type	Average number of players made per day	Average percentage of faulty players per day
Video players	2000	5%
Audio players	6000	3%

Question 2: FAULTY PLAYERS

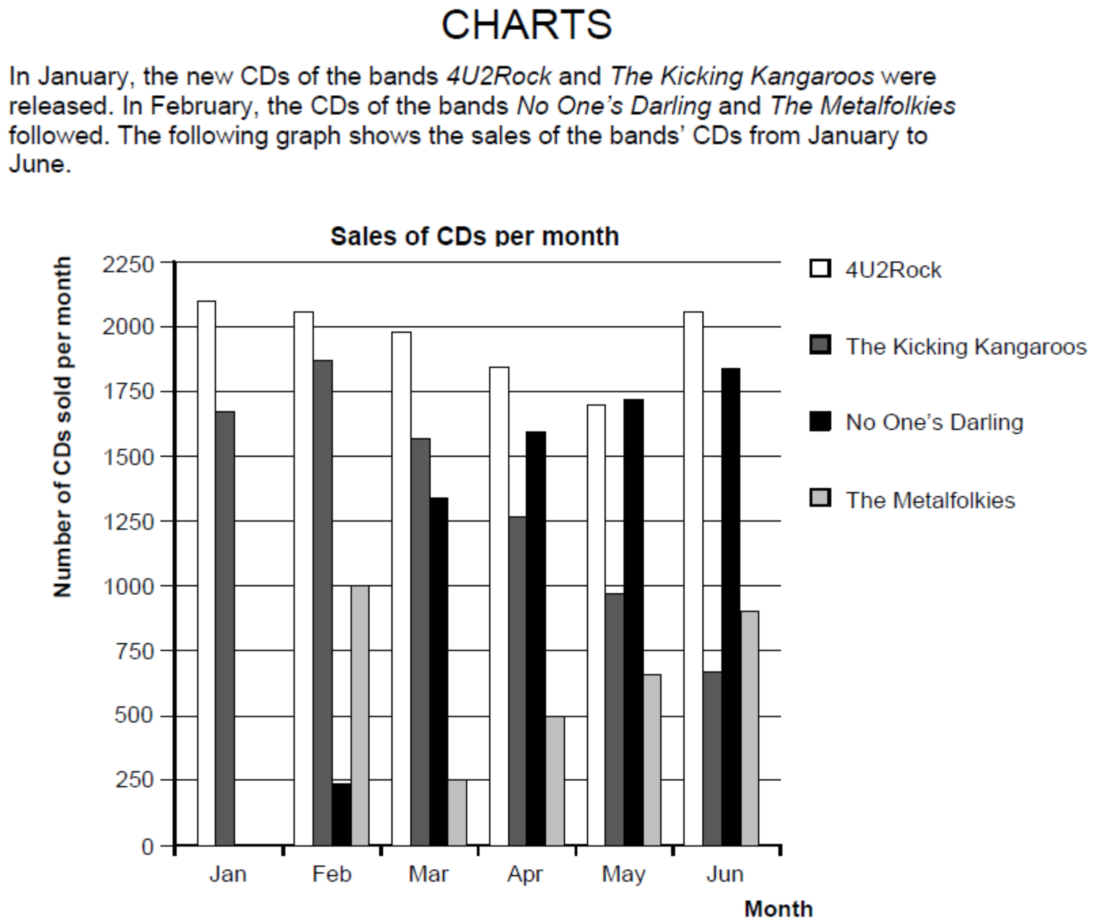
PM00EQ02 – 0 1 9

One of the testers makes the following claim:

"On average, there are more video players sent for repair per day compared to the number of audio players sent for repair per day."

Decide whether or not the tester's claim is correct. Give a mathematical argument to support your answer.

Charts 1



CHARTS

In January, the new CDs of the bands *4U2Rock* and *The Kicking Kangaroos* were released. In February, the CDs of the bands *No One's Darling* and *The Metalfolkies* followed. The following graph shows the sales of the bands' CDs from January to June.

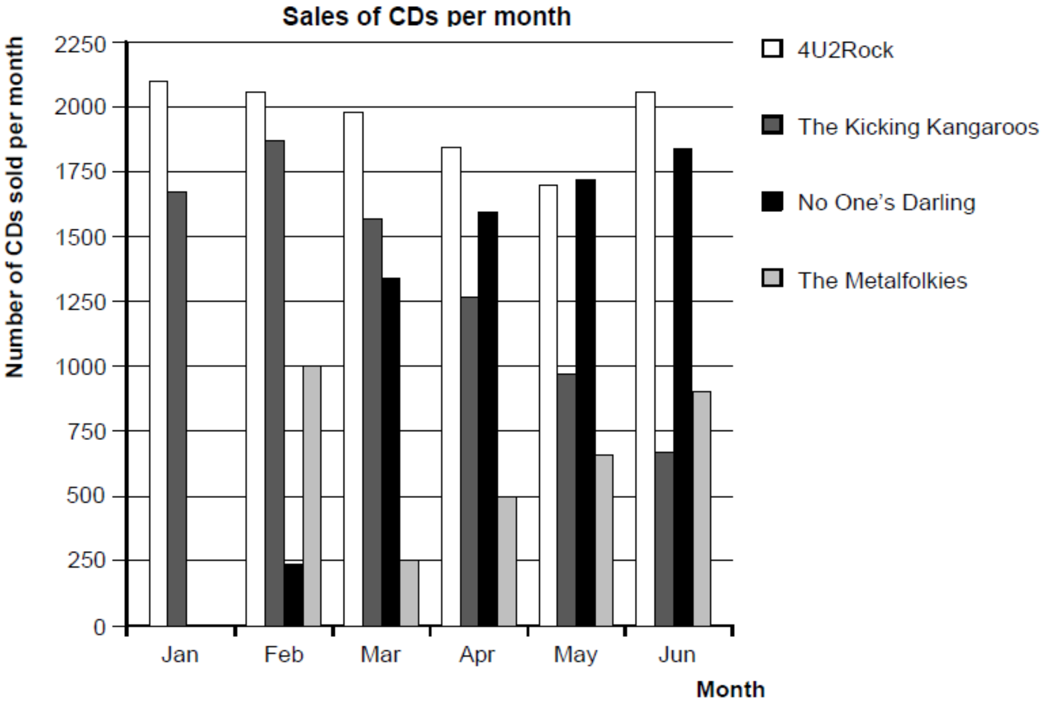
A question was created to assess students' communicating skill.

Please write a summary of important information from the graph.

Charts 2

CHARTS

In January, the new CDs of the bands *4U2Rock* and *The Kicking Kangaroos* were released. In February, the CDs of the bands *No One's Darling* and *The Metalfolkies* followed. The following graph shows the sales of the bands' CDs from January to June.

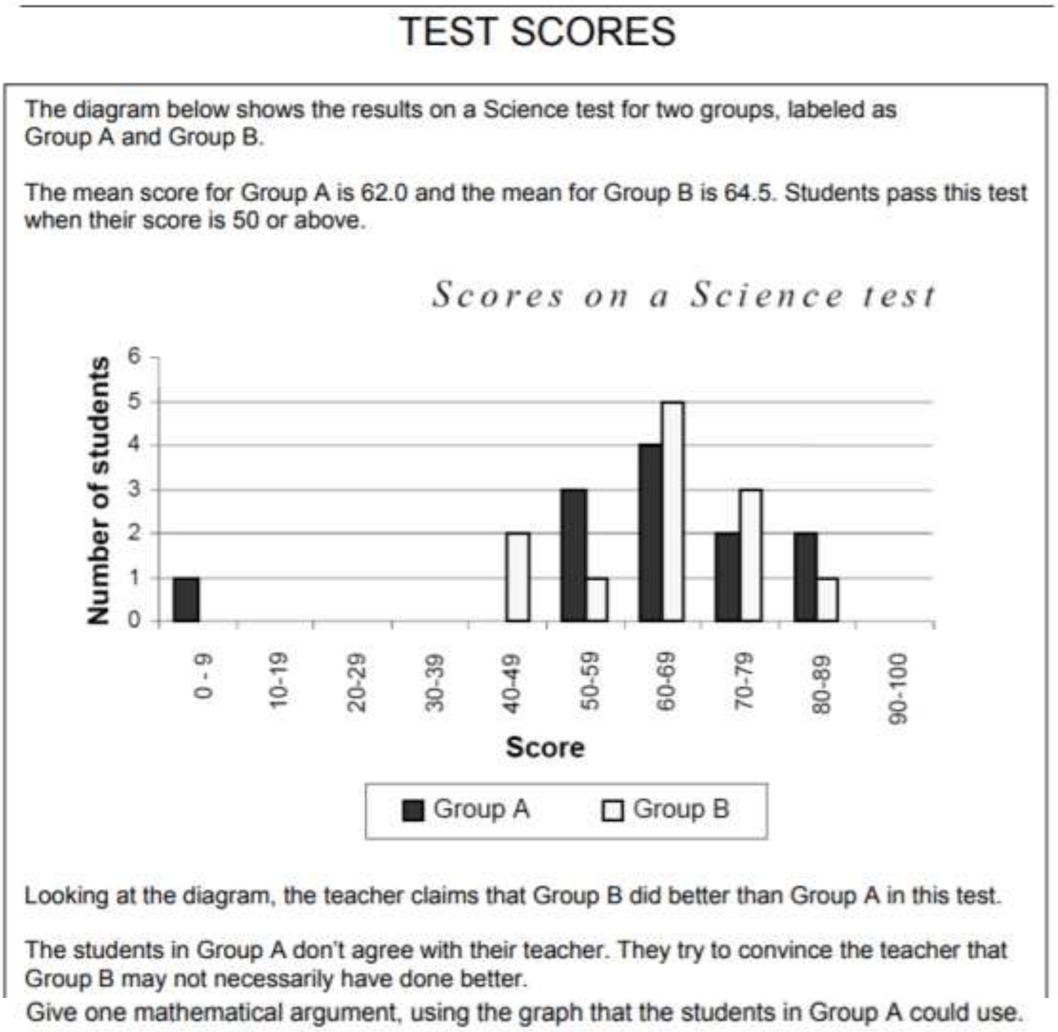


Question 2: CHARTS

PM918Q02

In which month did the band *No One's Darling* sell more CDs than the band *The Kicking Kangaroos* for the first time?

Test scores



TEST SCORES

The diagram below shows the results on a Science test for two groups, labeled as Group A and Group B.

The mean score for Group A is 62.0 and the mean for Group B is 64.5. Students pass this test when their score is 50 or above.

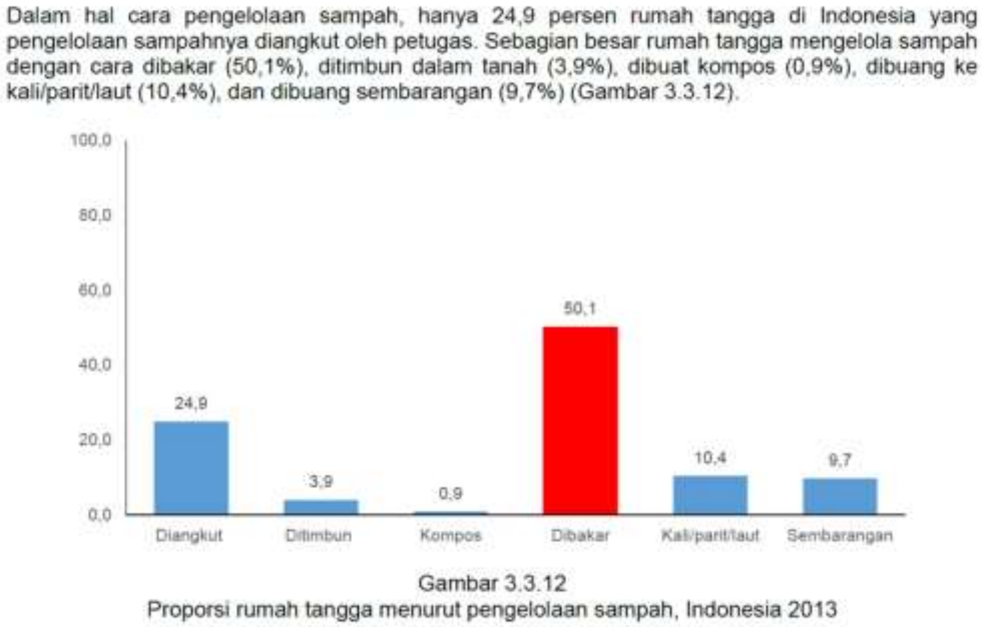
Scores on a Science test

Looking at the diagram, the teacher claims that Group B did better than Group A in this test.

The students in Group A don't agree with their teacher. They try to convince the teacher that Group B may not necessarily have done better.

Give one mathematical argument, using the graph that the students in Group A could use.

Domestic waste



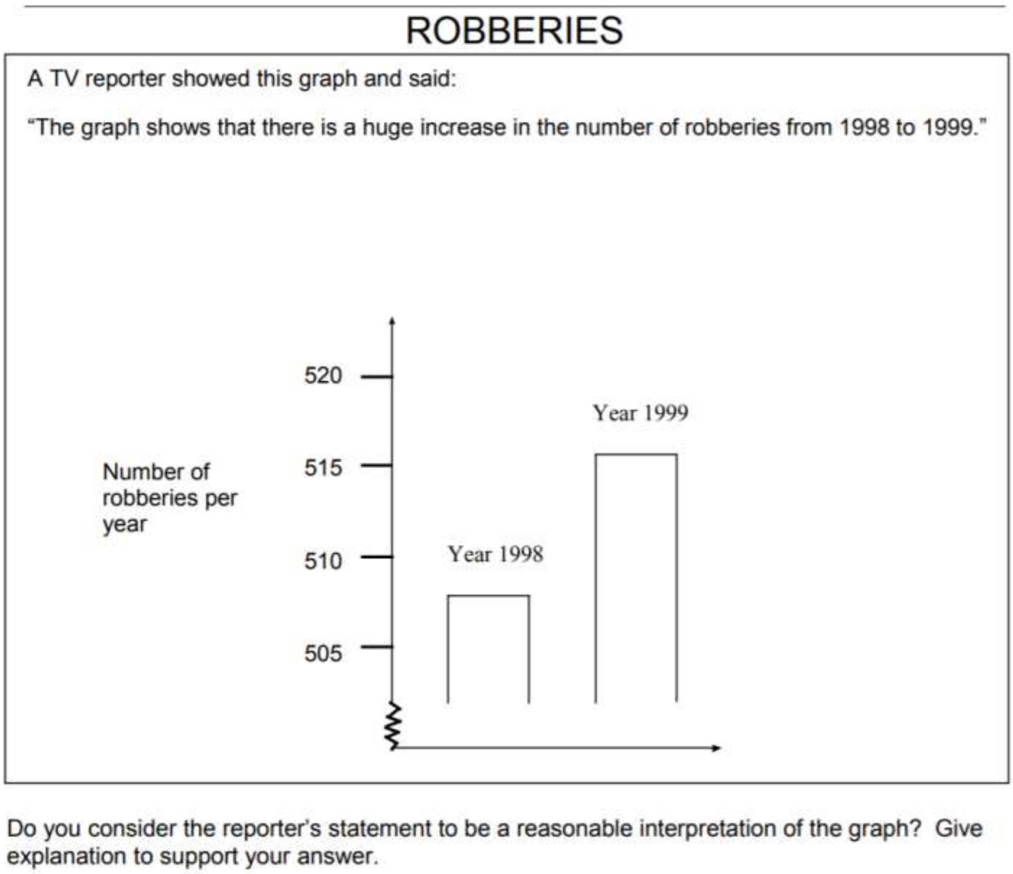
Dalam hal cara pengelolaan sampah, hanya 24,9 persen rumah tangga di Indonesia yang pengelolaan sampahnya diangkut oleh petugas. Sebagian besar rumah tangga mengelola sampah dengan cara dibakar (50,1%), ditimbun dalam tanah (3,9%), dibuat kompos (0,9%), dibuang ke kali/parit/laut (10,4%), dan dibuang sembarangan (9,7%) (Gambar 3.3.12).

Gambar 3.3.12
Proporsi rumah tangga menurut pengelolaan sampah, Indonesia 2013

A question would be created to assess students' communicating skill.

To make your friends informed, summaries the important information from the graph about the Indonesian people awareness of domestic waste management!

Robberies



ROBBERIES

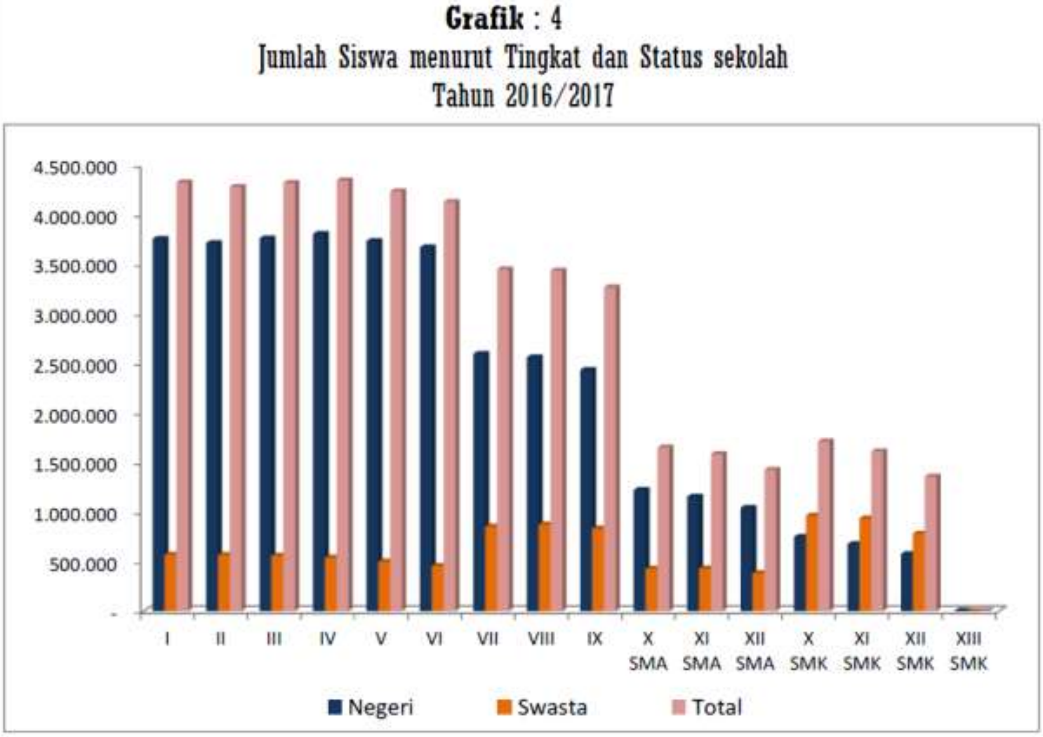
A TV reporter showed this graph and said:

"The graph shows that there is a huge increase in the number of robberies from 1998 to 1999."

520
515
510
505
Number of robberies per year
Year 1999
Year 1998

Do you consider the reporter's statement to be a reasonable interpretation of the graph? Give explanation to support your answer.

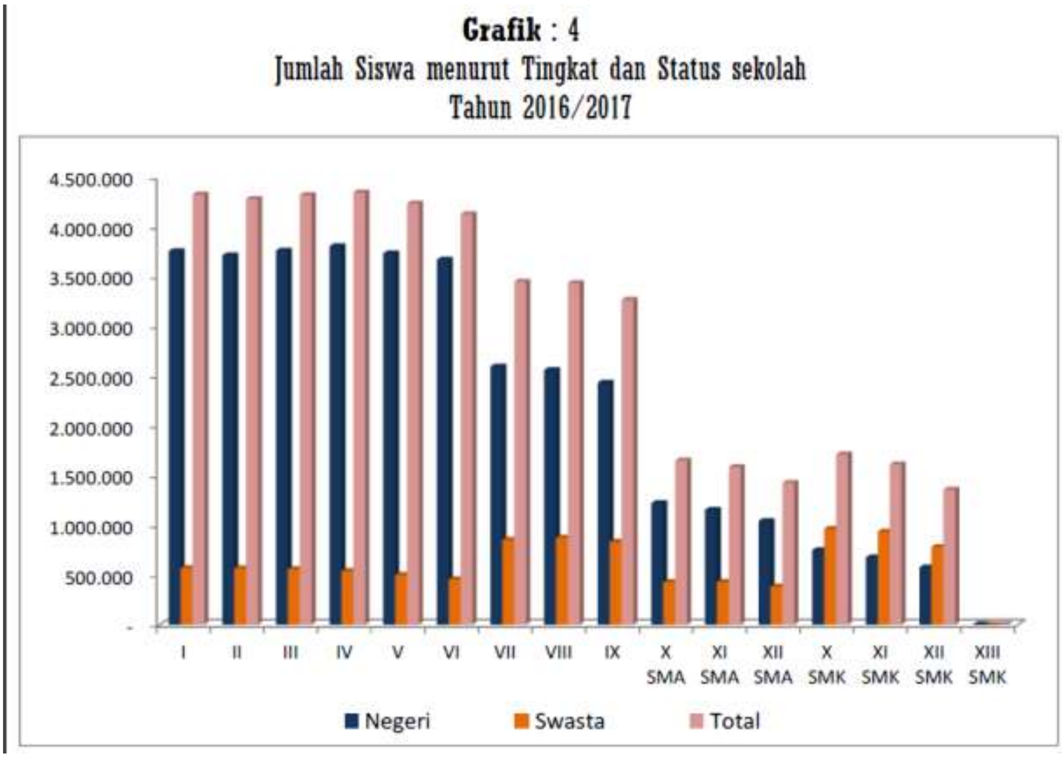
School students 1



A question was created to assess students' interpreting skill.

What are the numbers of students enrolled in Junior high school (VII to IX), and senior high school (X to XII)? Support your answer with supporting figures!

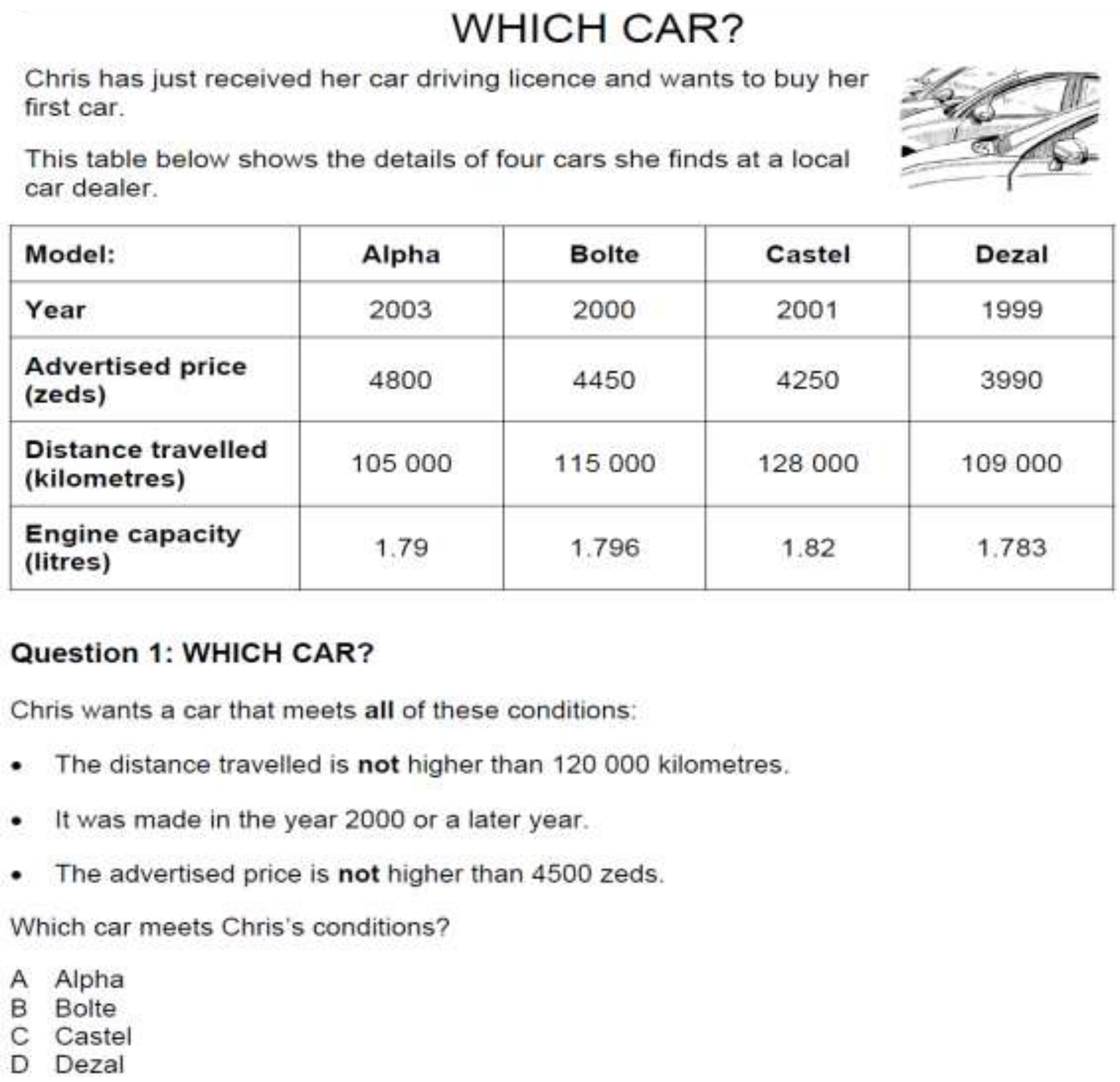
School students 2



A question was created to assess students' communicating skill.

Your friends conclude that there were dramatic decline on the number of students who continued to junior and senior high school levels. Give response to your friend's conclusion!

Which Car?



WHICH CAR?

Chris has just received her car driving licence and wants to buy her first car.

This table below shows the details of four cars she finds at a local car dealer.

Model:	**Alpha**	**Bolte**	**Castel**	**Dezal**
Year	2003	2000	2001	1999
Advertised price (zeds)	4800	4450	4250	3990
Distance travelled (kilometres)	105 000	115 000	128 000	109 000
Engine capacity (litres)	1.79	1.796	1.82	1.783

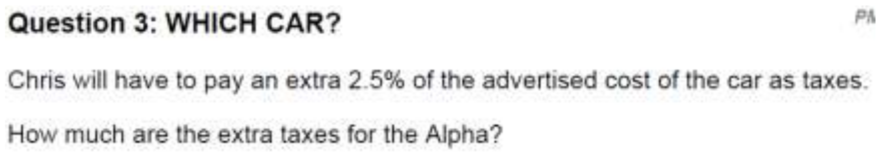
Question 1: WHICH CAR?

Chris wants a car that meets **all** of these conditions:

- The distance travelled is **not** higher than 120 000 kilometres.
- It was made in the year 2000 or a later year.
- The advertised price is **not** higher than 4500 zeds.

Which car meets Chris's conditions?

A Alpha
B Bolte
C Castel
D Dezal



Question 3: WHICH CAR?

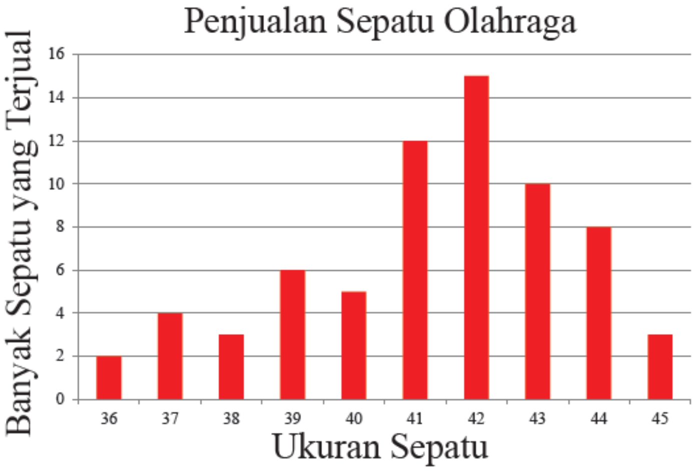
PA

Chris will have to pay an extra 2.5% of the advertised cost of the car as taxes.

How much are the extra taxes for the Alpha?

Sport shoes 1

Diagram berikut menunjukkan banyaknya sepatu olahraga yang terjual di Toko Sepatu Mantap Jaya pada bulan Agustus berdasarkan ukuran. Pemilik toko mengatakan bahwa sepatu olahraga yang terjual rata-rata adalah ukuran 42.

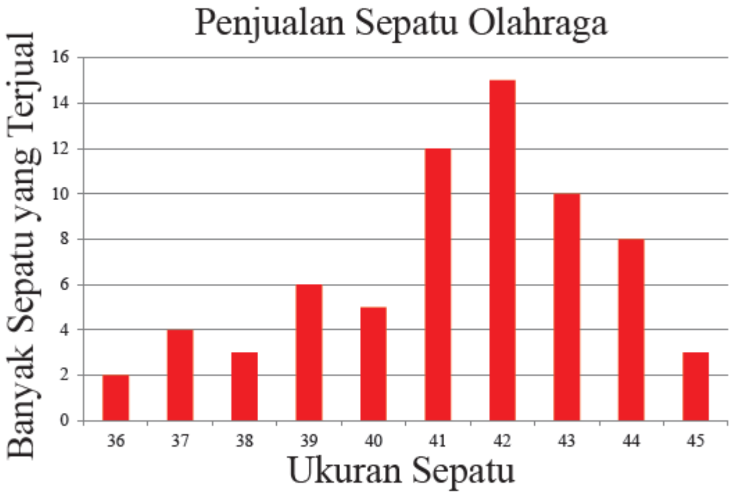


a. Dapatkan *mean*, median, dan modus dari data di atas. (untuk *mean* bulatkan sampai nilai satuan terdekat)

b. Apakah pernyataan pemilik toko tersebut benar? Jika salah, coba kamu betulkan pernyataan pemilik toko tersebut.

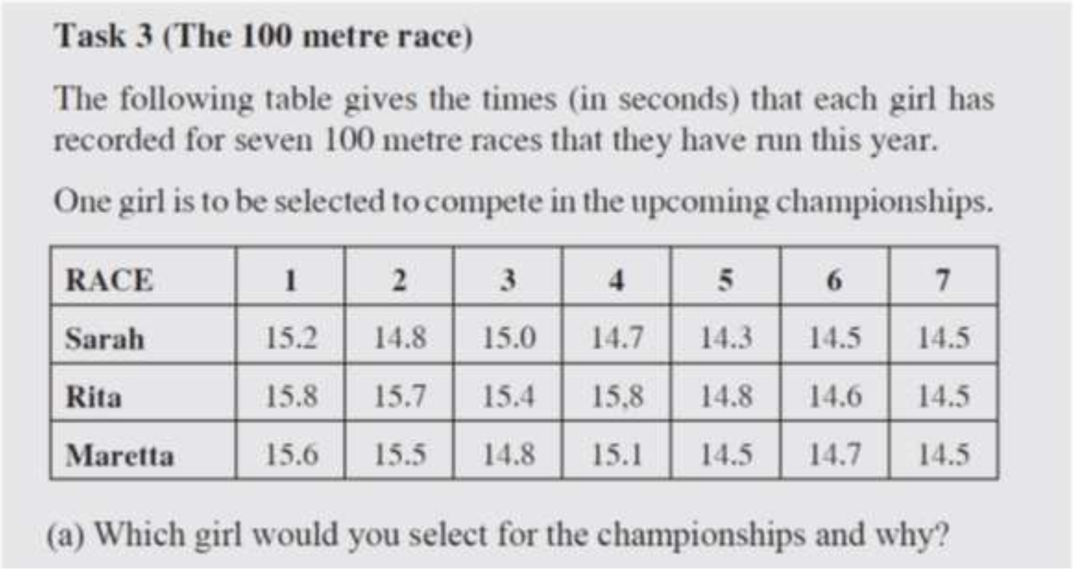
Sport shoes 2

Diagram berikut menunjukkan banyaknya sepatu olahraga yang terjual di Toko Sepatu Mantap Jaya pada bulan Agustus berdasarkan ukuran. Pemilik toko mengatakan bahwa sepatu olahraga yang terjual rata-rata adalah ukuran 42.



c. Pada bulan September, pemilik toko ingin menambah stok sepatu olahraga ukuran tertentu yang paling banyak terjual pada bulan sebelumnya, akan tetapi ia belum dapat menentukannya. Dengan menggunakan hasil yang telah kamu dapatkan pada poin a, perhitungan manakah yang dapat membantu pemilik toko dalam menyelesaikan permasalahan tersebut? Apakah *mean*, median, atau modus? Jelaskan jawabanmu.

The 100-metre race



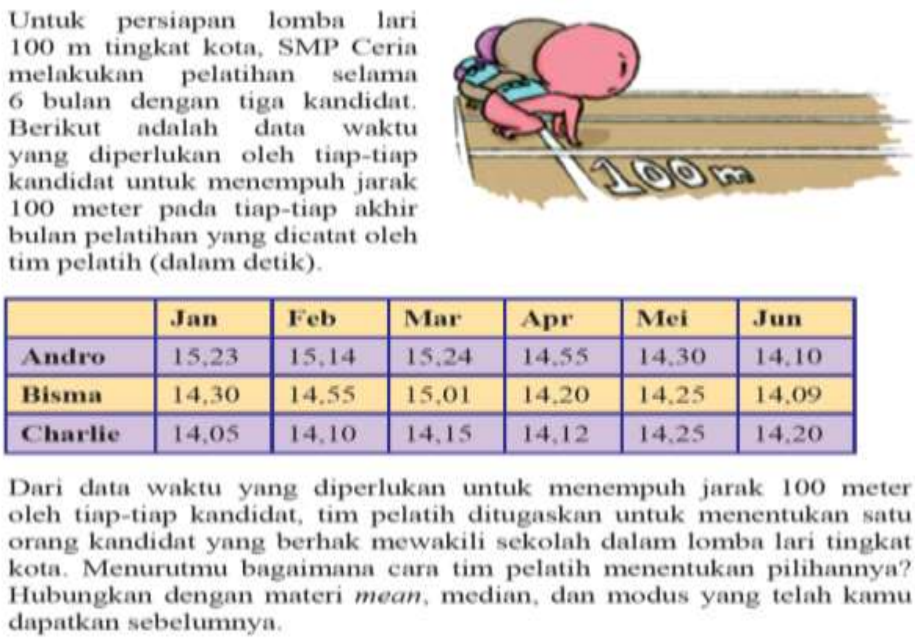
Task 3 (The 100 metre race)

The following table gives the times (in seconds) that each girl has recorded for seven 100 metre races that they have run this year.

One girl is to be selected to compete in the upcoming championships.

RACE	1	2	3	4	5	6	7
Sarah	15.2	14.8	15.0	14.7	14.3	14.5	14.5
Rita	15.8	15.7	15.4	15,8	14.8	14.6	14.5
Maretta	15.6	15.5	14.8	15.1	14.5	14.7	14.5

(a) Which girl would you select for the championships and why?

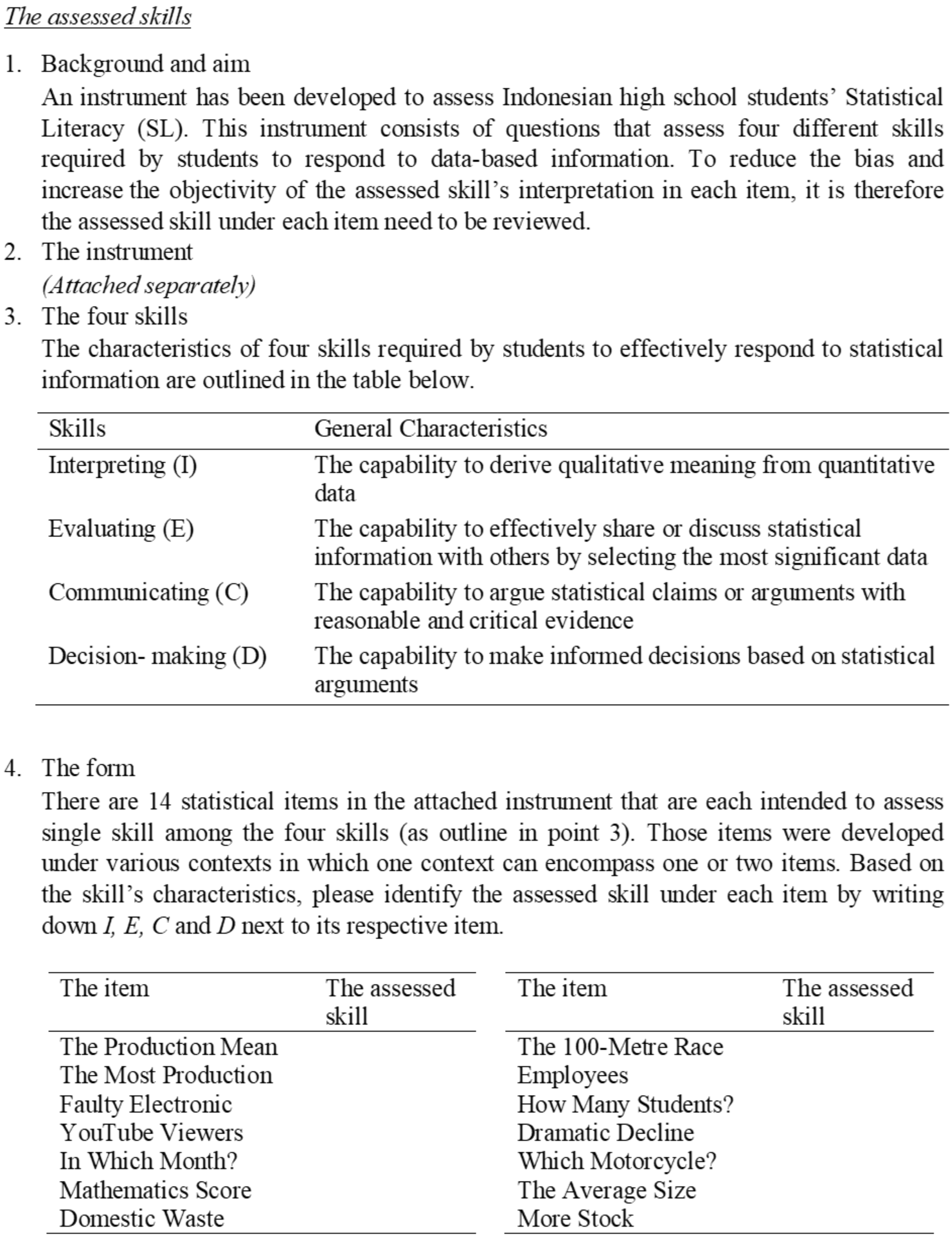


Untuk persiapan lomba lari 100 m tingkat kota, SMP Ceria melakukan pelatihan selama 6 bulan dengan tiga kandidat. Berikut adalah data waktu yang diperlukan oleh tiap-tiap kandidat untuk menempuh jarak 100 meter pada tiap-tiap akhir bulan pelatihan yang dicatat oleh tim pelatih (dalam detik).

	Jan	Feb	Mar	Apr	Mei	Jun
Andro	15,23	15,14	15,24	14,55	14,30	14,10
Bisma	14,30	14,55	15,01	14,20	14,25	14,09
Charlie	14,05	14,10	14,15	14,12	14,25	14,20

Dari data waktu yang diperlukan untuk menempuh jarak 100 meter oleh tiap-tiap kandidat, tim pelatih ditugaskan untuk menentukan satu orang kandidat yang berhak mewakili sekolah dalam lomba lari tingkat kota. Menurutmu bagaimana cara tim pelatih menentukan pilihannya? Hubungkan dengan materi *mean*, median, dan modus yang telah kamu dapatkan sebelumnya.

Appendix B. Skill Review Form



The assessed skills

1. Background and aim

   An instrument has been developed to assess Indonesian high school students' Statistical Literacy (SL). This instrument consists of questions that assess four different skills required by students to respond to data-based information. To reduce the bias and increase the objectivity of the assessed skill's interpretation in each item, it is therefore the assessed skill under each item need to be reviewed.
2. The instrument

   (Attached separately)
3. The four skills

   The characteristics of four skills required by students to effectively respond to statistical information are outlined in the table below.

Skills	General Characteristics
Interpreting (I)	The capability to derive qualitative meaning from quantitative data
Evaluating (E)	The capability to effectively share or discuss statistical information with others by selecting the most significant data
Communicating (C)	The capability to argue statistical claims or arguments with reasonable and critical evidence
Decision- making (D)	The capability to make informed decisions based on statistical arguments

4. The form

   There are 14 statistical items in the attached instrument that are each intended to assess single skill among the four skills (as outline in point 3). Those items were developed under various contexts in which one context can encompass one or two items. Based on the skill's characteristics, please identify the assessed skill under each item by writing down *I, E, C* and *D* next to its respective item.

The item	The assessed skill	The item	The assessed skill
The Production Mean		The 100-Metre Race	
The Most Production		Employees	
Faulty Electronic		How Many Students?	
YouTube Viewers		Dramatic Decline	
In Which Month?		Which Motorcycle?	
Mathematics Score		The Average Size	
Domestic Waste		More Stock	

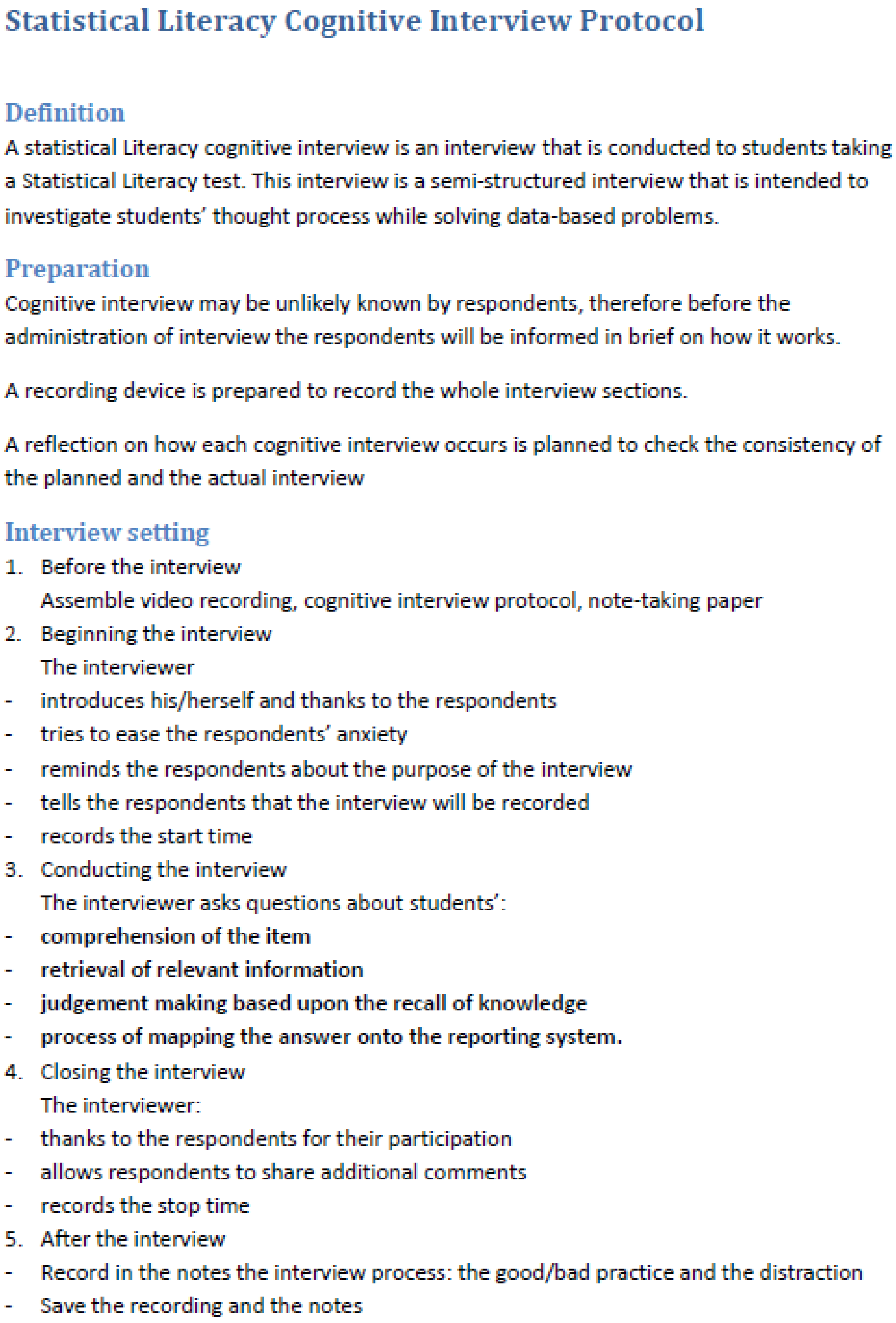
5. Additional inquiries

Lastly, the researcher is seeking your help to suggest him which items you think better to include in the test. Please tick (√) on the selected items, otherwise cross (×) it.

The item	Tick (√) or cross (×)
The Production Mean	
The Most Production	
Faulty Electronic	
YouTube Viewers	
In Which Month?	
Mathematics Score	
Domestic Waste	

The item	Tick (√) or cross (×)
The 100-Metre Race	
Employees	
How Many Students?	
Dramatic Decline	
Which Motorcycle?	
The Average Size	
More Stock	

Appendix C. Initial Interview Protocol



Statistical Literacy Cognitive Interview Protocol

Definition

A statistical Literacy cognitive interview is an interview that is conducted to students taking a Statistical Literacy test. This interview is a semi-structured interview that is intended to investigate students' thought process while solving data-based problems.

Preparation

Cognitive interview may be unlikely known by respondents, therefore before the administration of interview the respondents will be informed in brief on how it works.

A recording device is prepared to record the whole interview sections.

A reflection on how each cognitive interview occurs is planned to check the consistency of the planned and the actual interview

Interview setting

1. Before the interview

   Assemble video recording, cognitive interview protocol, note-taking paper
2. Beginning the interview

   The interviewer
- introduces his/herself and thanks to the respondents
- tries to ease the respondents' anxiety
- reminds the respondents about the purpose of the interview
- tells the respondents that the interview will be recorded
- records the start time

3. Conducting the interview

   The interviewer asks questions about students':
- **comprehension of the item**
- **retrieval of relevant information**
- **judgement making based upon the recall of knowledge**
- **process of mapping the answer onto the reporting system.**

4. Closing the interview

   The interviewer:
- thanks to the respondents for their participation
- allows respondents to share additional comments
- records the stop time

5. After the interview
- Record in the notes the interview process: the good/bad practice and the distraction
- Save the recording and the notes

Interview script (general)

1. Beginning the interview (4 minutes)

Before we begin this interview, let's have a proper introduction. My name is Badrun, and you are .... Nice to meet you.
First of all, I would like to say thank you very much for participating in this interview. Your participation will help me to understand what is going on in your mind while you are working with data-based problems. There will be no right and wrong answers; therefore, you don't have to be afraid of. Any information you provide during this interview will be recorded. Is that OK? .... Thank you.

During this interview, please think aloud as you're solving the problems. It means say anything what you think; I'm interested in hearing all your thought and reactions. ***(repeat and emphasize this information).***

We will now start this cognitive interview at time ....

2. Conducting the interview (25 minutes)

- The comprehension of the item
    - *Please start solving this item by reading its texts aloud!*
    - **Probe:** *Do you understand what you've just read?*

If No, Probe:	*Which particular information in this item that you hardly understand?*
If Yes, Probe:	*Can you explain it in brief?*

-
    - *That's great. Thinking out loud like this is just what I need*

- The retrieval of relevant information
    - *I am interested in what you are thinking as you retrieve relevant information from the problem?, do whatever you need to help you think aloud.*
    - **Probe:** *Why do you think that might be relevant information?*
    - *Thank you, your reaction is really helpful*

- The judgement making based upon the recall of knowledge
    - *Do you understand what this question is asking for?*
    - **Probe:** *Then, what's all you need to solve it?*
    - *That's fine you are talking through your reaction and it is very helpful for me*

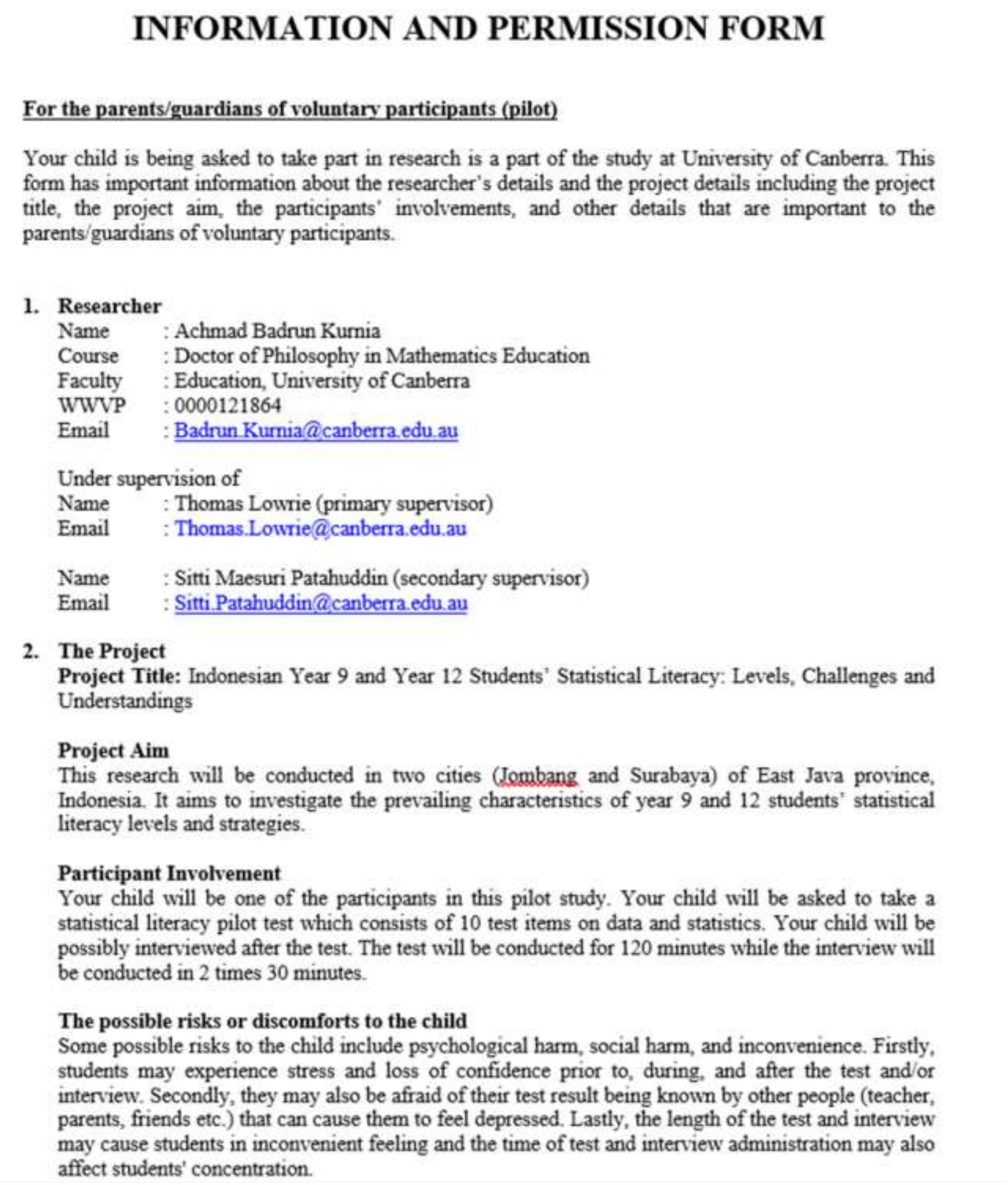
- The process of mapping the answer onto the reporting system.
    - *Please start writing down the answer for this question* and *do whatever you need by saying it out loud.*

(this series of questions applies for all item test)

3. Closing the interview (1 minute)

Thank you for taking time to participate in this interview. If you have any comments to share, please feel free. ....
That's the end of this interview, and now I will stop the recording at time ......

Appendix D. Sample of Information Sheet and Consent Form for Parents



INFORMATION AND PERMISSION FORM

For the parents/guardians of voluntary participants (pilot)

Your child is being asked to take part in research is a part of the study at University of Canberra. This form has important information about the researcher's details and the project details including the project title, the project aim, the participants' involvements, and other details that are important to the parents/guardians of voluntary participants.

1. **Researcher**

Name : Achmad Badrun Kurnia
Course : Doctor of Philosophy in Mathematics Education
Faculty : Education, University of Canberra
WWVP : 0000121864
Email : Badrun.Kurnia@canberra.edu.au

Under supervision of
Name : Thomas Lowrie (primary supervisor)
Email : Thomas.Lowrie@canberra.edu.au

Name : Sitti Maesuri Patahuddin (secondary supervisor)
Email : Sitti.Patahuddin@canberra.edu.au

2. **The Project**

Project Title: Indonesian Year 9 and Year 12 Students' Statistical Literacy: Levels, Challenges and Understandings

Project Aim

This research will be conducted in two cities (Jombang and Surabaya) of East Java province, Indonesia. It aims to investigate the prevailing characteristics of year 9 and 12 students' statistical literacy levels and strategies.

Participant Involvement

Your child will be one of the participants in this pilot study. Your child will be asked to take a statistical literacy pilot test which consists of 10 test items on data and statistics. Your child will be possibly interviewed after the test. The test will be conducted for 120 minutes while the interview will be conducted in 2 times 30 minutes.

The possible risks or discomforts to the child

Some possible risks to the child include psychological harm, social harm, and inconvenience. Firstly, students may experience stress and loss of confidence prior to, during, and after the test and/or interview. Secondly, they may also be afraid of their test result being known by other people (teacher, parents, friends etc.) that can cause them to feel depressed. Lastly, the length of the test and interview may cause students in inconvenient feeling and the time of test and interview administration may also affect students' concentration.

However, to minimize this potential risks, the researcher (with the help of math teachers) will anticipate this situation by informing that: (1) the test will not affect their mathematics performances at school, (2) the aim for the pilot test/interview is not for grading but for helping researcher' study, (3) the result of the test will be confidential and not given back to the students, schools, or parents, (4) only researcher that are able to track the students' real databases, (5) the pilot test/interview will be conducted beyond school time, (6) interview will be conducted at maximum 2 times 30 minutes, and (7) snack and refreshment will be provided during the pilot test.

Confidentiality

Results of this study may be used in thesis report as well as publications and presentations. It will be guaranteed that the information from your child will be handled with respect and discretion. The test result and personal information attached to it will not be available for general consumption. Only the researcher has access to your individual data. Data is collected only for the stated research purpose, and will not be used for any other purposes without your permission. To ensure and safeguard your data, identifying information will be removed. Participants' individual data will be anonymized in the thesis or any publication arising from this research.

Anonymity

The information collected from participants will be non-identifiable. Pseudonyms and codes will be used for analysis of the data. Only the researcher will know the code.

Data Storage

All records of data will be kept secure by the researcher. Following University of Canberra protocols, data will be stored for a five year period, after which it will be destroyed.

Data usage

Data obtained from this study may be used in researcher's future research project for publications and presentations after the completion his candidature.

Ethics Committee Clearance

This research has been approved by the University of Canberra Ethics Committee in Human Research of the University of Canberra.

Financial Information

Participation in this study will involve no cost to you or your child. Your child will not be paid for participating in this study.

The child's rights as a research participant

Participation in this study is voluntary. Your child may withdraw from this study at any time.You and your child will not be penalized in any way or lose any sort of benefits for deciding to stop participation. If you and your child decide not to be in this study, this will not affect the relationship you and your child have with your child's school in any way. Your child's grades will not be affected if you choose not to let your child be in this study.

If your child decides to withdraw from this study, the researchers will ask if the information already collected from your child can be used.

Queries and Concerns

For further information or questions concerning this project, please contact the researcher on 0435949022 or Badrun.Kurnia@canberra.edu.au

Parental Permission for Child's Participation in Research

I have read this form thoroughly and been given the opportunity to ask questions. If I have additional questions, I have been told whom to contact. Therefore, I give permission for my child to participate in the research study described above and will receive a copy of this Parental Permission form after I sign it.

____ ____________

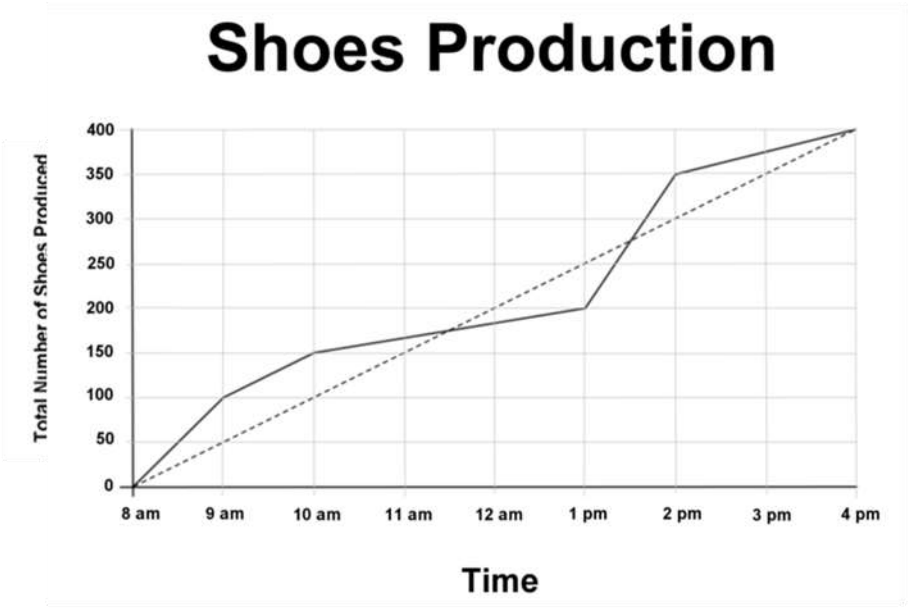
Parent/Legal Guardian's Name (printed) and Signature Date

____ ____________

Name of Person Obtaining Parental Permission Date

Appendix E. Revised Items after Pilot Interview I

Shoe production



The solid line (——) on graph shows shoes production by a home industry during a particular day.

The dotted line (- - - -) shows what the total number of shoes produced would be if the rate of production were constant

During which hour were the most cars produced? Show your steps to find it!

Faulty electronic

The *Toshiba Company* makes two types of electronic devices: TV and computer. At the end of the daily production, both TV and computer are tested and those with faults are removed and sent for repair.

The following table shows the average number of each of the two electronic devices that are made per day, and the average percentage of faulty devices per day.

Electronic device type	**Average number of devices made per day**	**Average percentage of faulty devices per day**
TV	2000	5%
Computer	6000	3%

One of the testers makes the following claim:

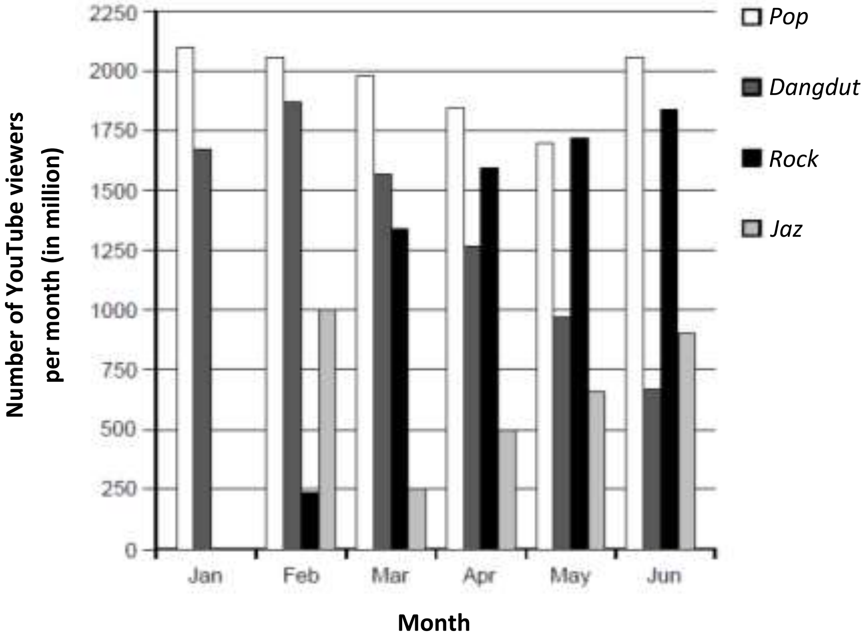
“On average, there are more TV sent for repair per day compared to the number of computer sent for repair per day”

Is the tester’s claim correct? Show your process to support/argue it.

YouTube viewers

In January, the new single of the bands *Pop* and *Dangdut* were released. In February, the single of the bands *Rock* and *Jaz* followed. The following graph shows the number of their YouTube viewers from January to June.

YouTube viewers per month



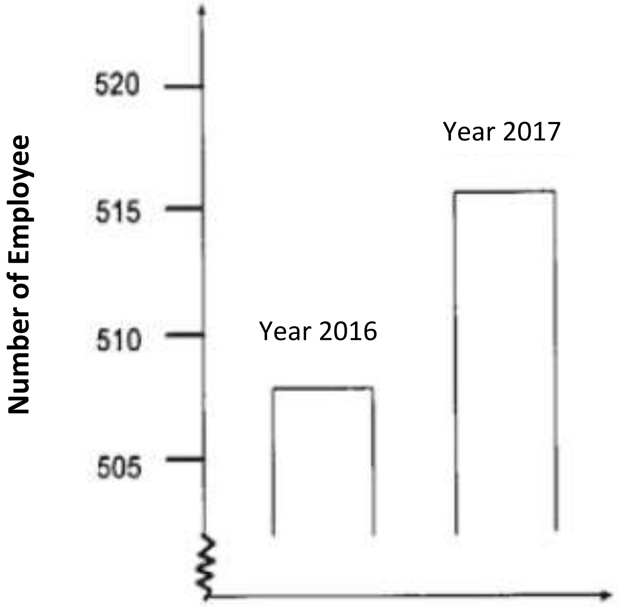
Please write a summary of important information from the graph?

In which month did the number of viewers for Rock band single exceed that for Dangdut band single for the first time? Use estimate to show the difference in the number of viewers for both bands at that time.

Employees

A newspaper reader read this graph and said:

“The graph shows that there is a huge increase in number of employee from 2016 to 2017”



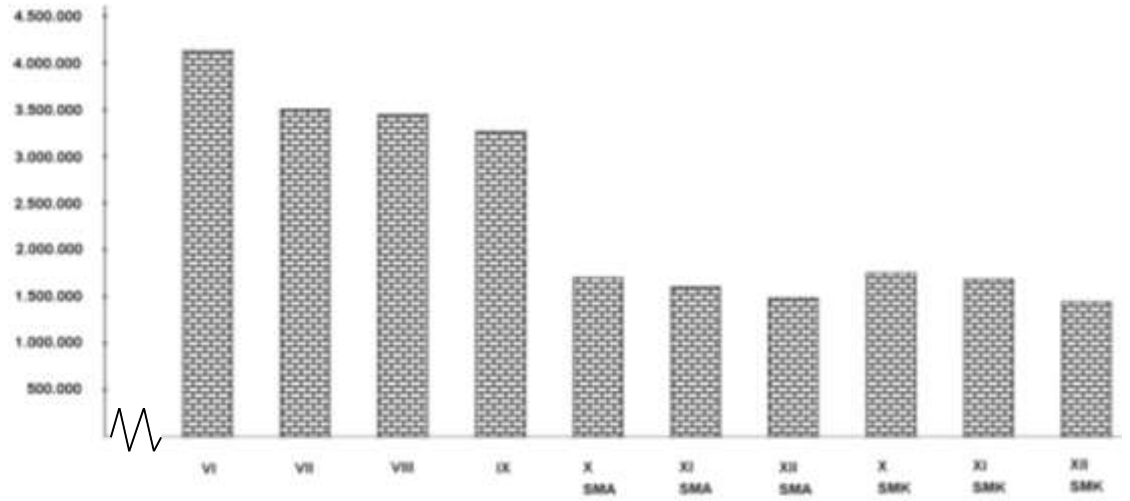
Do you consider the reader’s statement to be a reasonable interpretation of the graph?

Give an explanation to support your answer.

School students

The number of Indonesian students (grade 6 to 12)

in 2016/2017



What are the numbers of students enrolled in Junior High School and Senior High School? Support your answer with supporting figures!

Your friends conclude that there were dramatic decline on the number of students who continued to junior high school levels. Give response to your friend’s conclusion!

Which motorcycle?
Rano wants to buy a second-hand motorcycle that meets all of these conditions:

- The distance travelled is not higher than 35,000 kilometres
- It was made in the year 2010 or a later year
- The advertised price is not higher than Rp 6,500,000

He decides to go to a local dealer and he finds the motorcycles' details as shown in the table below.

Model:	**Jupiter A**	**Jupiter B**	**Jupiter C**	**Jupiter D**
Year	2015	2012	2013	2011
Advertised price* (thousands Rupiah)	6,800	6,450	6,250	5,990
Distance travelled (kilometres)	29,000	34,000	35,000	32,000

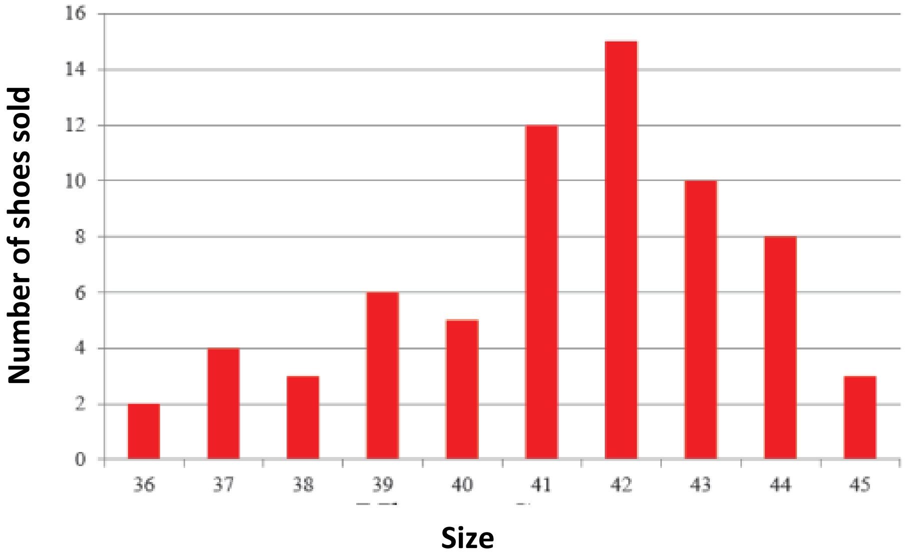
*Excluding extra cost (taxes) 2.5%

Which motorcycle is best for Rano? Show your steps and reasoning to choose the motorcycle!

Sport shoes

The graph below shows the number of sport shoes that is sold in *Mantap Jaya* shop on August based on the size.

The sales of sport shoes



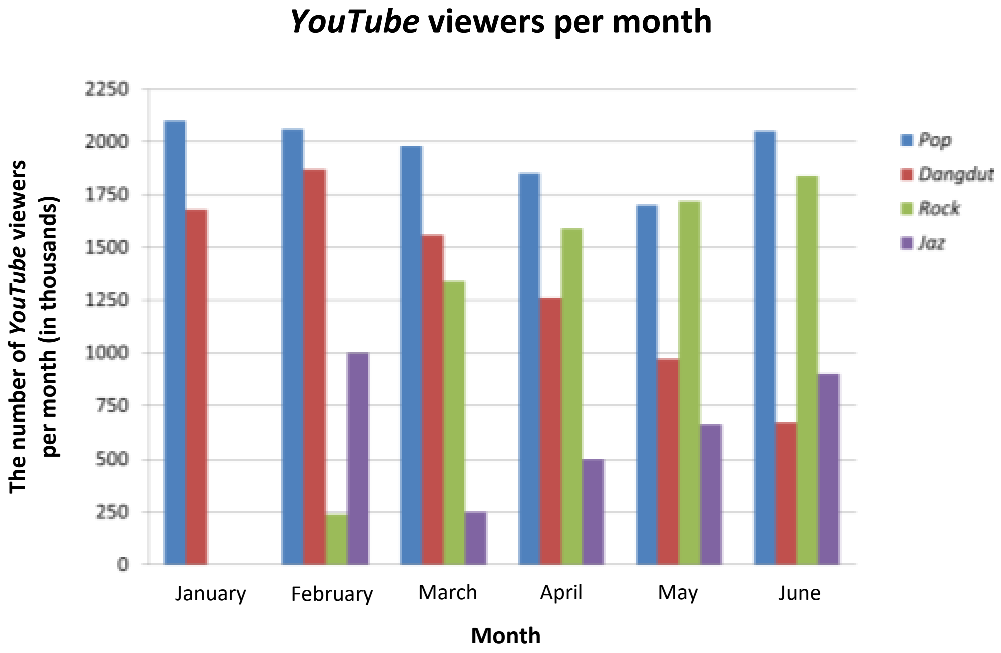
The shop owner claim that on average the sold out shoes are size 42. Is this claim correct? Give your argument!

Next month the shop owner wants to add his collection on certain size that was mostly sold in August. However, the shop owner cannot decide what size he needs to add more. How can you help him to decide?

Appendix F. Sample of Test Items

Among the ten items, below are the two items assessing interpreting skills, while the other eight items were explained in Section 4.4.

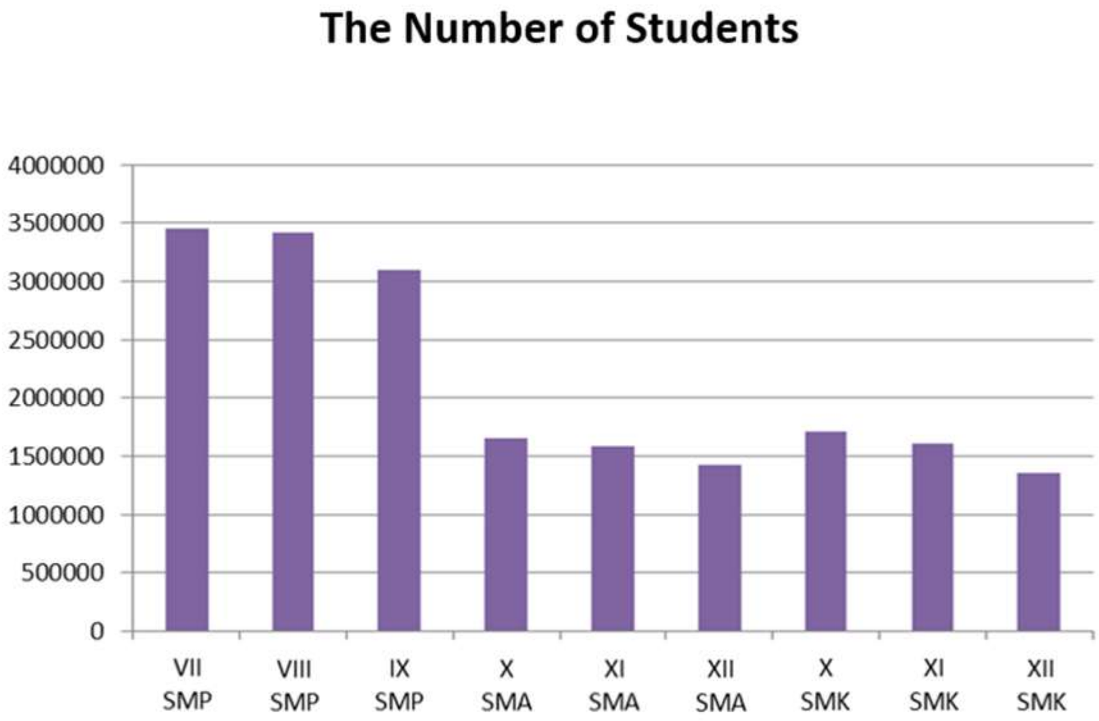
In January, new singles by the bands *Pop* and *Dangdut* were released. In February, singles by the bands *Rock* and *Jazz* followed. The following graph shows these bands' number of *YouTube* viewers from January to June.



In which month did the number of viewers for Rock band single exceed that for Dangdut band single for the first time? Use estimate to show the difference in the number of viewers for both bands at that time.

School Students

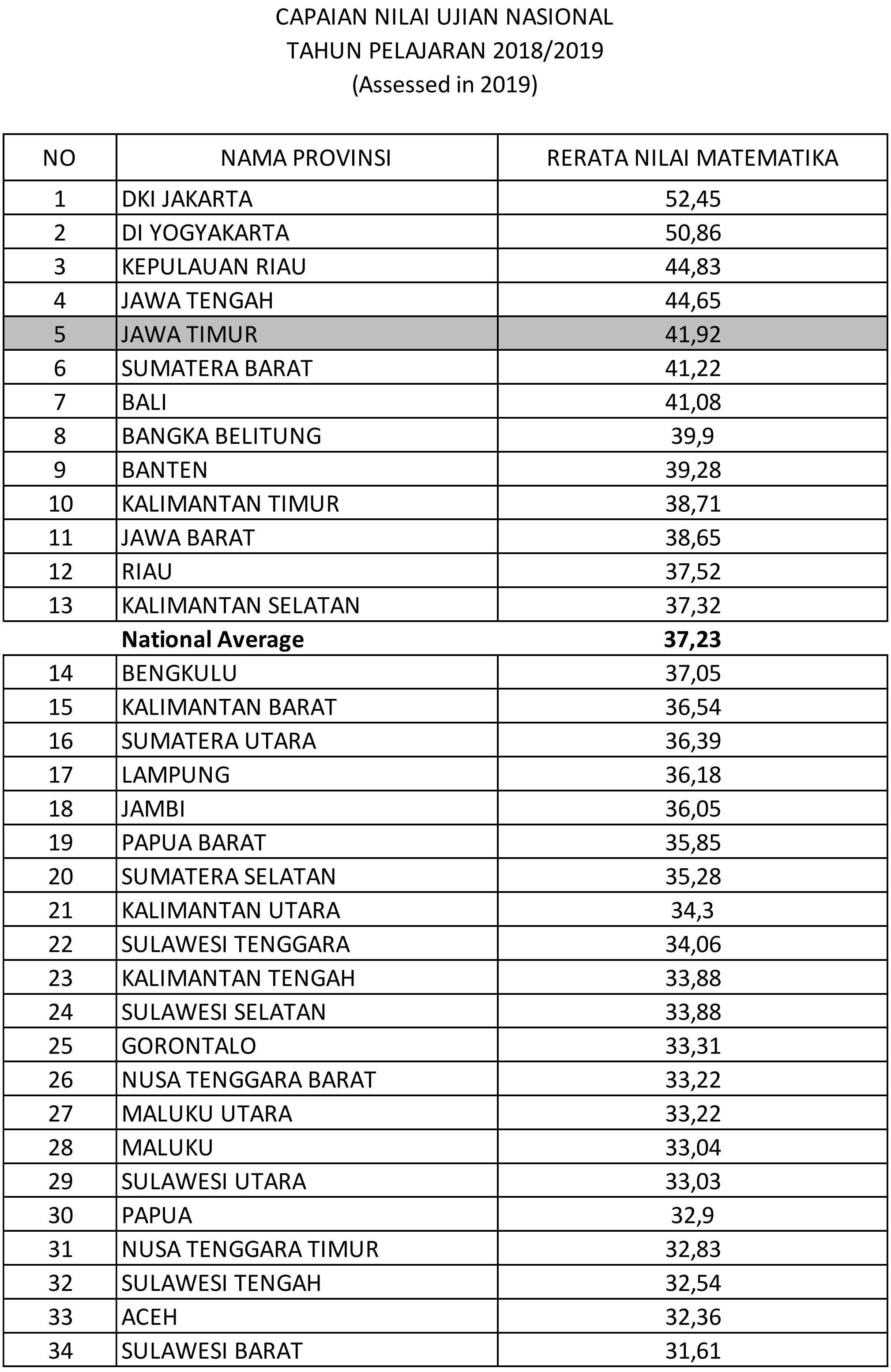
The Ministry of Education and Culture gathered information, which is shown in the bar graph below, to determine the number of Indonesian students enrolled in Junior High School (*SMP*), Senior High School (*SMA*) and Vocational School (*SMK*) in 2016–2017.



Give your opinion about the number of students in IX SMP and X SMA in the bar graph above? Use the estimates to support your answer.

Appendix G. Average Mathematics Score of UN 2019 for All Provinces in Indonesia

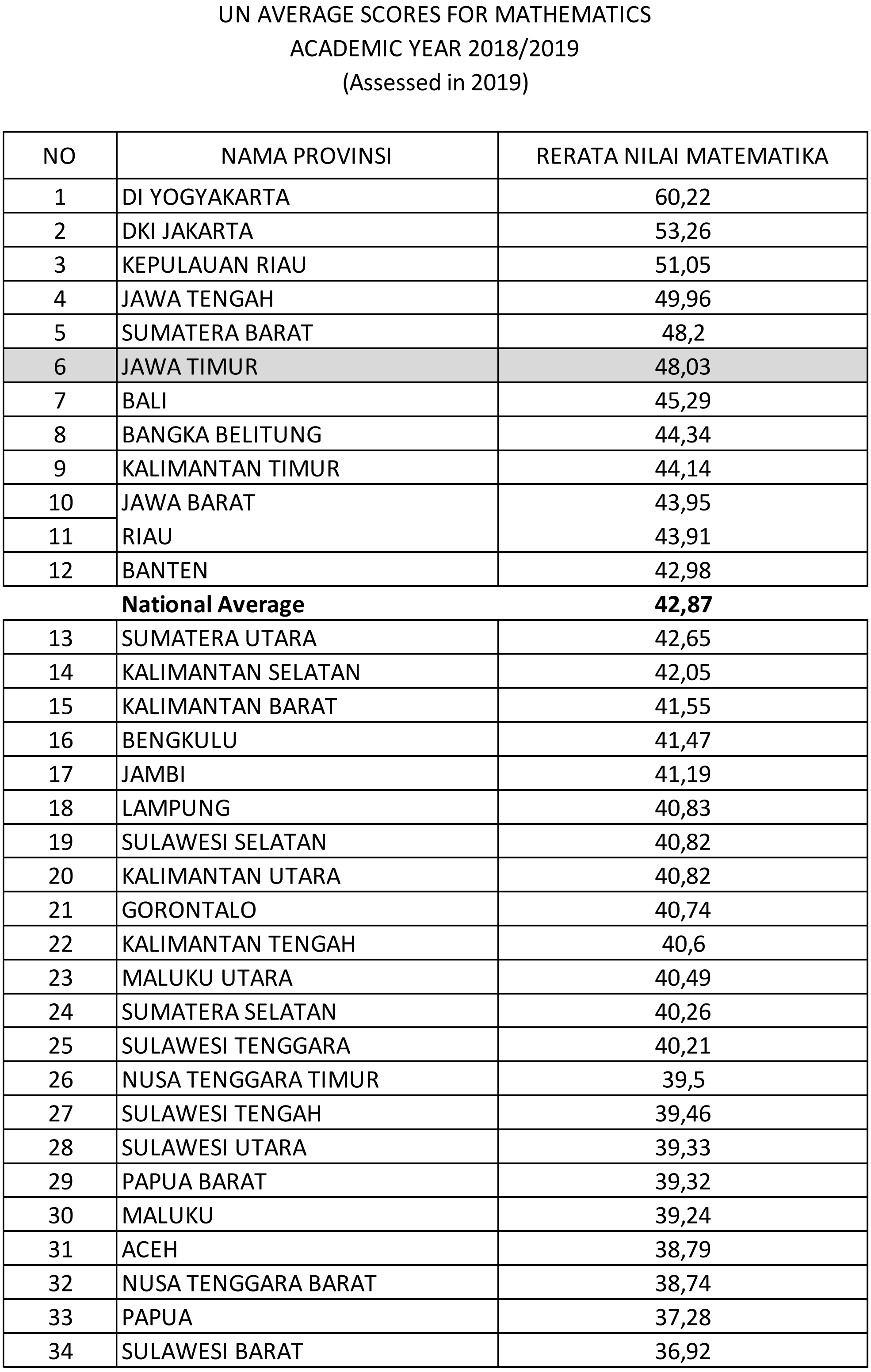
Table G. 1 *Average Score for Senior High School Students by Provinces*



CAPAIAN NILAI UJIAN NASIONAL
TAHUN PELAJARAN 2018/2019
(Assessed in 2019)

NO	NAMA PROVINSI	RERATA NILAI MATEMATIKA
1	DKI JAKARTA	52,45
2	DI YOGYAKARTA	50,86
3	KEPULAUAN RIAU	44,83
4	JAWA TENGAH	44,65
5	JAWA TIMUR	41,92
6	SUMATERA BARAT	41,22
7	BALI	41,08
8	BANGKA BELITUNG	39,9
9	BANTEN	39,28
10	KALIMANTAN TIMUR	38,71
11	JAWA BARAT	38,65
12	RIAU	37,52
13	KALIMANTAN SELATAN	37,32
	National Average	**37,23**
14	BENGKULU	37,05
15	KALIMANTAN BARAT	36,54
16	SUMATERA UTARA	36,39
17	LAMPUNG	36,18
18	JAMBI	36,05
19	PAPUA BARAT	35,85
20	SUMATERA SELATAN	35,28
21	KALIMANTAN UTARA	34,3
22	SULAWESI TENGGARA	34,06
23	KALIMANTAN TENGAH	33,88
24	SULAWESI SELATAN	33,88
25	GORONTALO	33,31
26	NUSA TENGGARA BARAT	33,22
27	MALUKU UTARA	33,22
28	MALUKU	33,04
29	SULAWESI UTARA	33,03
30	PAPUA	32,9
31	NUSA TENGGARA TIMUR	32,83
32	SULAWESI TENGAH	32,54
33	ACEH	32,36
34	SULAWESI BARAT	31,61

Table G. 2 *Average Score for Junior High School Students by Provinces*

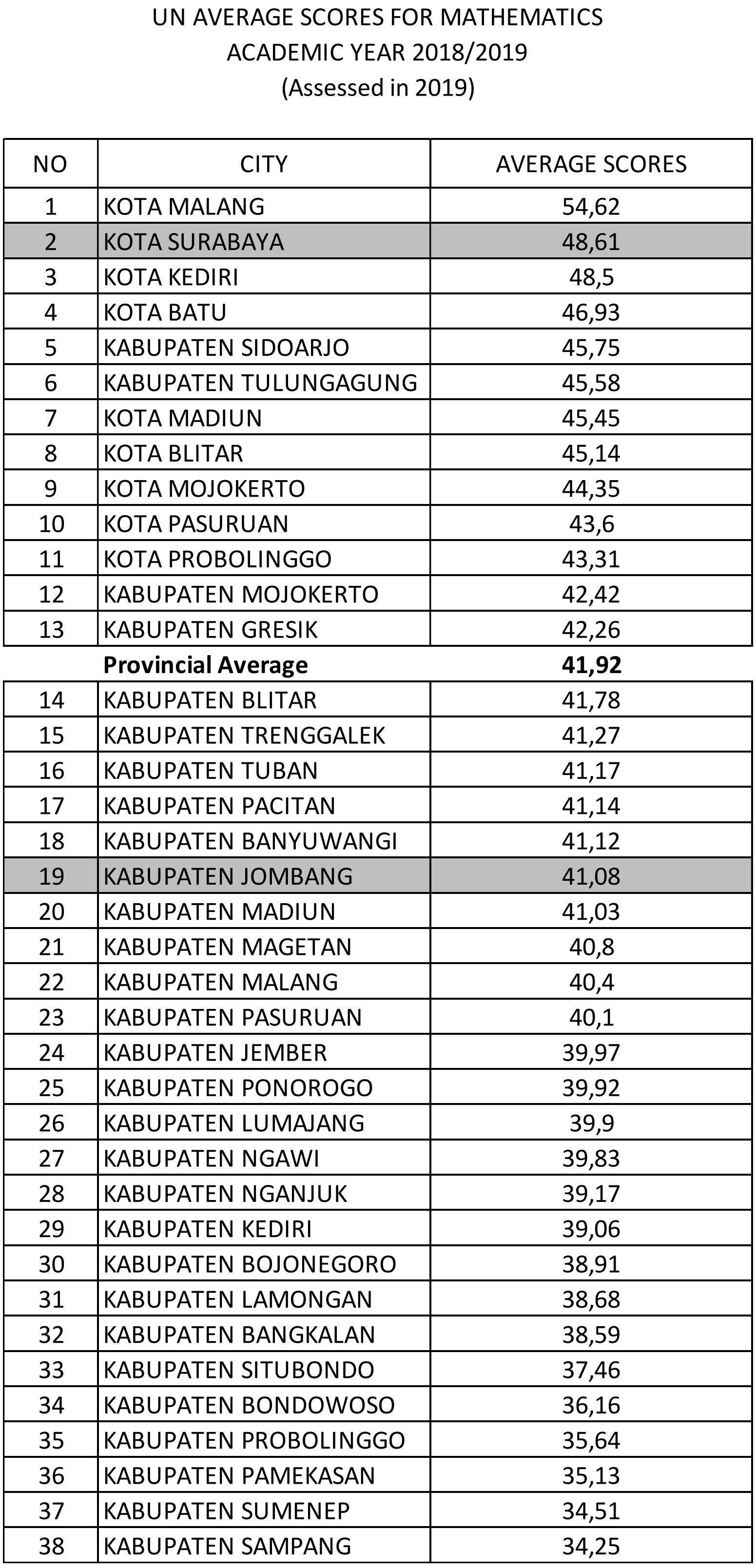


UN AVERAGE SCORES FOR MATHEMATICS
ACADEMIC YEAR 2018/2019
(Assessed in 2019)

NO	NAMA PROVINSI	RERATA NILAI MATEMATIKA
1	DI YOGYAKARTA	60,22
2	DKI JAKARTA	53,26
3	KEPULAUAN RIAU	51,05
4	JAWA TENGAH	49,96
5	SUMATERA BARAT	48,2
6	JAWA TIMUR	48,03
7	BALI	45,29
8	BANGKA BELITUNG	44,34
9	KALIMANTAN TIMUR	44,14
10	JAWA BARAT	43,95
11	RIAU	43,91
12	BANTEN	42,98
	National Average	**42,87**
13	SUMATERA UTARA	42,65
14	KALIMANTAN SELATAN	42,05
15	KALIMANTAN BARAT	41,55
16	BENGKULU	41,47
17	JAMBI	41,19
18	LAMPUNG	40,83
19	SULAWESI SELATAN	40,82
20	KALIMANTAN UTARA	40,82
21	GORONTALO	40,74
22	KALIMANTAN TENGAH	40,6
23	MALUKU UTARA	40,49
24	SUMATERA SELATAN	40,26
25	SULAWESI TENGGARA	40,21
26	NUSA TENGGARA TIMUR	39,5
27	SULAWESI TENGAH	39,46
28	SULAWESI UTARA	39,33
29	PAPUA BARAT	39,32
30	MALUKU	39,24
31	ACEH	38,79
32	NUSA TENGGARA BARAT	38,74
33	PAPUA	37,28
34	SULAWESI BARAT	36,92

Appendix H. Average Mathematics Score of UN 2019 for All Cities in East Java

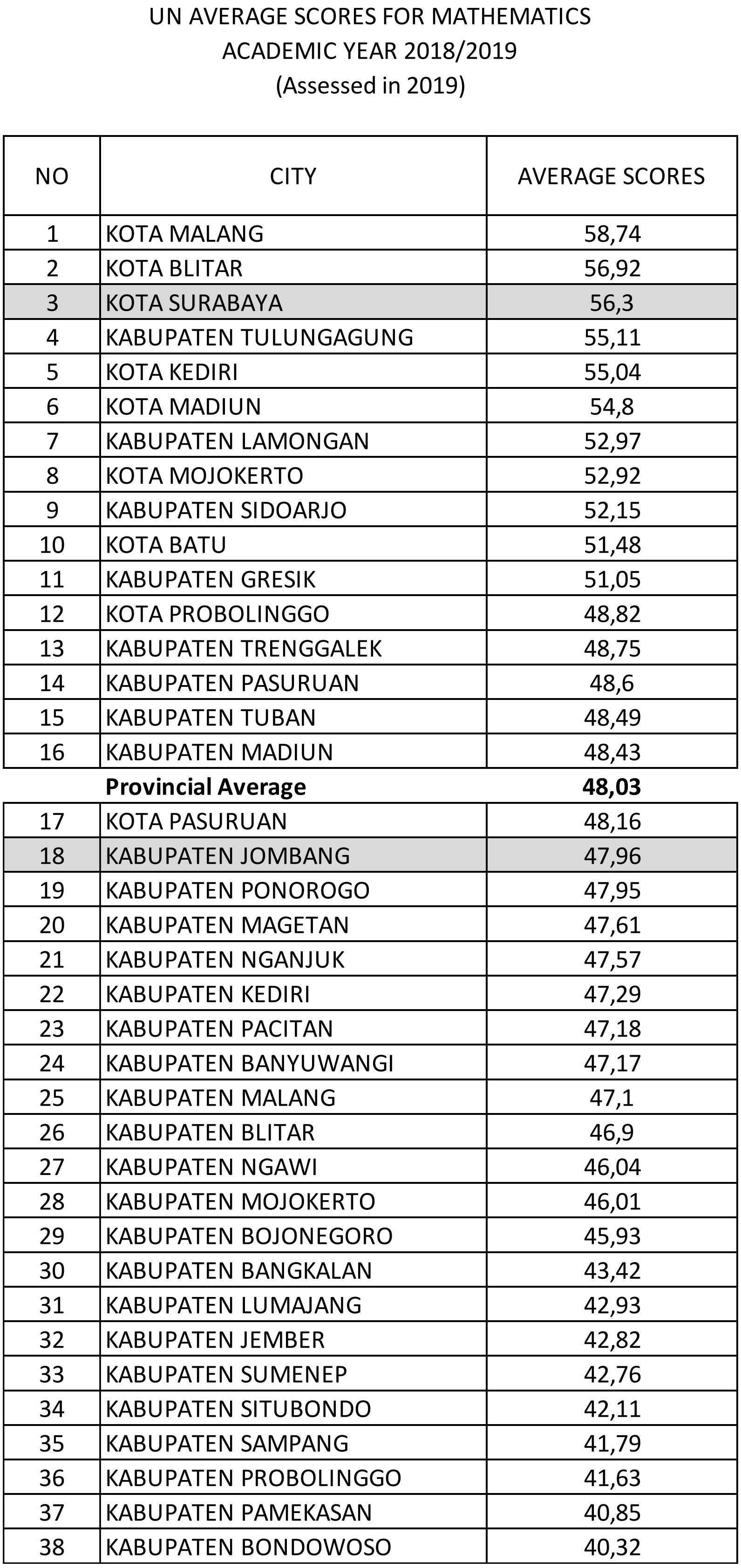
Table H. 1 *Average Score for Senior High School Students by Cities in East Java*



UN AVERAGE SCORES FOR MATHEMATICS
ACADEMIC YEAR 2018/2019
(Assessed in 2019)

NO	CITY	AVERAGE SCORES
1	KOTA MALANG	54,62
2	KOTA SURABAYA	48,61
3	KOTA KEDIRI	48,5
4	KOTA BATU	46,93
5	KABUPATEN SIDOARJO	45,75
6	KABUPATEN TULUNGAGUNG	45,58
7	KOTA MADIUN	45,45
8	KOTA BLITAR	45,14
9	KOTA MOJOKERTO	44,35
10	KOTA PASURUAN	43,6
11	KOTA PROBOLINGGO	43,31
12	KABUPATEN MOJOKERTO	42,42
13	KABUPATEN GRESIK	42,26
	Provincial Average	**41,92**
14	KABUPATEN BLITAR	41,78
15	KABUPATEN TRENGGALEK	41,27
16	KABUPATEN TUBAN	41,17
17	KABUPATEN PACITAN	41,14
18	KABUPATEN BANYUWANGI	41,12
19	KABUPATEN JOMBANG	41,08
20	KABUPATEN MADIUN	41,03
21	KABUPATEN MAGETAN	40,8
22	KABUPATEN MALANG	40,4
23	KABUPATEN PASURUAN	40,1
24	KABUPATEN JEMBER	39,97
25	KABUPATEN PONOROGO	39,92
26	KABUPATEN LUMAJANG	39,9
27	KABUPATEN NGAWI	39,83
28	KABUPATEN NGANJUK	39,17
29	KABUPATEN KEDIRI	39,06
30	KABUPATEN BOJONEGORO	38,91
31	KABUPATEN LAMONGAN	38,68
32	KABUPATEN BANGKALAN	38,59
33	KABUPATEN SITUBONDO	37,46
34	KABUPATEN BONDOWOSO	36,16
35	KABUPATEN PROBOLINGGO	35,64
36	KABUPATEN PAMEKASAN	35,13
37	KABUPATEN SUMENEP	34,51
38	KABUPATEN SAMPANG	34,25

Table H. 2 *Average Score for Junior High School Students by Cities in East Java*



UN AVERAGE SCORES FOR MATHEMATICS
ACADEMIC YEAR 2018/2019
(Assessed in 2019)

NO	CITY	AVERAGE SCORES
1	KOTA MALANG	58,74
2	KOTA BLITAR	56,92
3	KOTA SURABAYA	56,3
4	KABUPATEN TULUNGAGUNG	55,11
5	KOTA KEDIRI	55,04
6	KOTA MADIUN	54,8
7	KABUPATEN LAMONGAN	52,97
8	KOTA MOJOKERTO	52,92
9	KABUPATEN SIDOARJO	52,15
10	KOTA BATU	51,48
11	KABUPATEN GRESIK	51,05
12	KOTA PROBOLINGGO	48,82
13	KABUPATEN TRENGGALEK	48,75
14	KABUPATEN PASURUAN	48,6
15	KABUPATEN TUBAN	48,49
16	KABUPATEN MADIUN	48,43
	Provincial Average	**48,03**
17	KOTA PASURUAN	48,16
18	KABUPATEN JOMBANG	47,96
19	KABUPATEN PONOROGO	47,95
20	KABUPATEN MAGETAN	47,61
21	KABUPATEN NGANJUK	47,57
22	KABUPATEN KEDIRI	47,29
23	KABUPATEN PACITAN	47,18
24	KABUPATEN BANYUWANGI	47,17
25	KABUPATEN MALANG	47,1
26	KABUPATEN BLITAR	46,9
27	KABUPATEN NGAWI	46,04
28	KABUPATEN MOJOKERTO	46,01
29	KABUPATEN BOJONEGORO	45,93
30	KABUPATEN BANGKALAN	43,42
31	KABUPATEN LUMAJANG	42,93
32	KABUPATEN JEMBER	42,82
33	KABUPATEN SUMENEP	42,76
34	KABUPATEN SITUBONDO	42,11
35	KABUPATEN SAMPANG	41,79
36	KABUPATEN PROBOLINGGO	41,63
37	KABUPATEN PAMEKASAN	40,85
38	KABUPATEN BONDOWOSO	40,32

Appendix I. Four-Character Code for All Tested Students

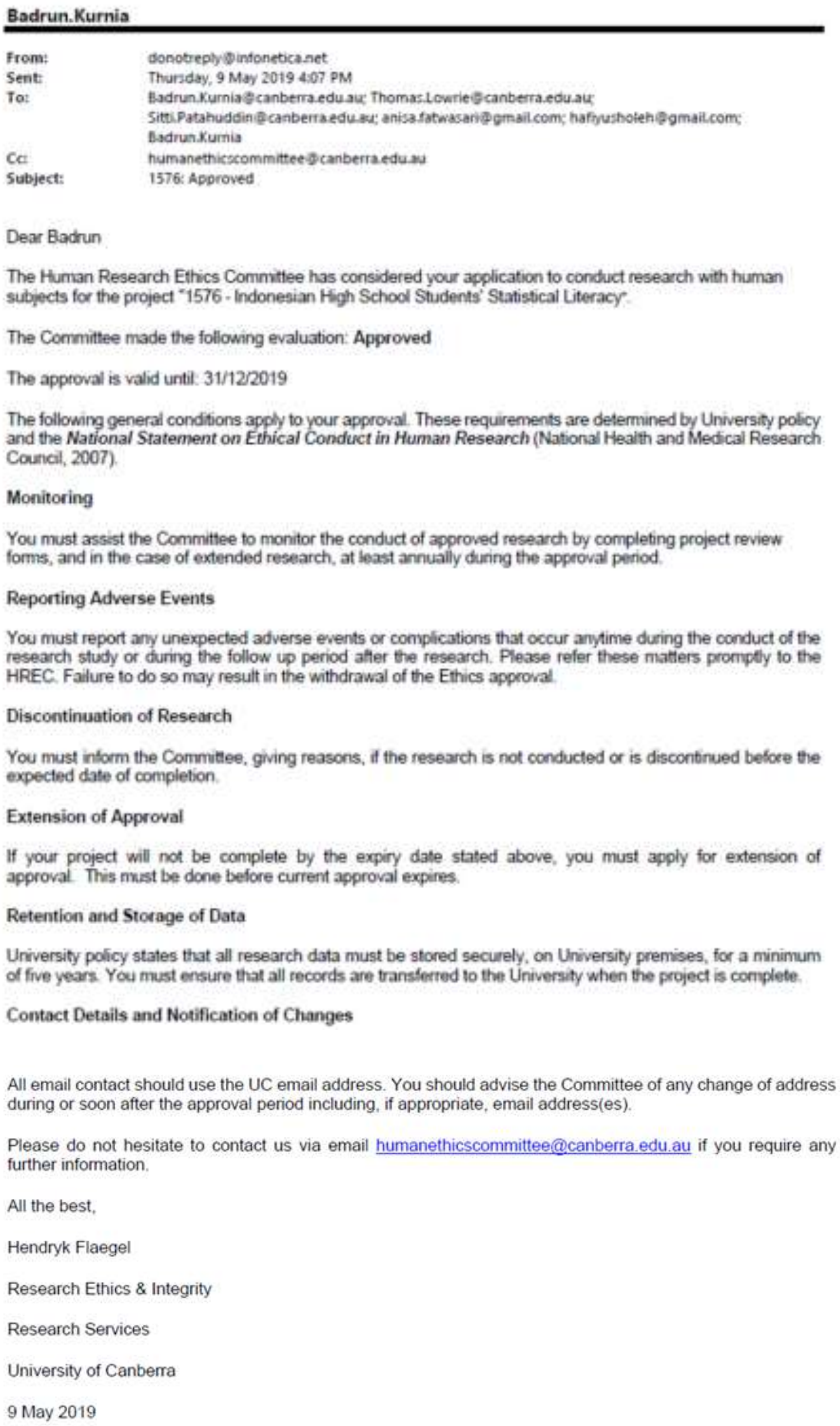
Table I. 1 *Four-Character Code for Senior High School Students*

No	Year	School type	School status	School	Gender	Student
1	Year 12	MoEC-RT	Public	Jombang	Girl	B01J
2	Year 12	MoEC-RT	Public	Jombang	Girl	B02J
3	Year 12	MoEC-RT	Public	Jombang	Girl	B03J
4	Year 12	MoEC-RT	Public	Jombang	Boy	B04J
5	Year 12	MoEC-RT	Public	Jombang	Boy	B05J
6	Year 12	MoEC-RT	Public	Jombang	Boy	B06J
7	Year 12	MoEC-RT	Private	Jombang	Girl	B07J
8	Year 12	MoEC-RT	Private	Jombang	Girl	B08J
9	Year 12	MoEC-RT	Private	Jombang	Girl	B09J
10	Year 12	MoEC-RT	Private	Jombang	Boy	B10J
11	Year 12	MoEC-RT	Private	Jombang	Boy	B11J
12	Year 12	MoEC-RT	Private	Jombang	Boy	B12J
13	Year 12	MoRA	Private	Jombang	Girl	B13J
14	Year 12	MoRA	Private	Jombang	Girl	B14J
15	Year 12	MoRA	Private	Jombang	Girl	B15J
16	Year 12	MoRA	Private	Jombang	Boy	B16J
17	Year 12	MoRA	Private	Jombang	Boy	B17J
18	Year 12	MoRA	Private	Jombang	Boy	B18J
19	Year 12	MoRA	Public	Jombang	Girl	B19J
20	Year 12	MoRA	Public	Jombang	Girl	B20J
21	Year 12	MoRA	Public	Jombang	Girl	B21J
22	Year 12	MoRA	Public	Jombang	Boy	B22J
23	Year 12	MoRA	Public	Jombang	Boy	B23J
24	Year 12	MoRA	Public	Jombang	Boy	B24J
25	Year 12	MoEC-RT	Public	Surabaya	Girl	B01S
26	Year 12	MoEC-RT	Public	Surabaya	Girl	B02S
27	Year 12	MoEC-RT	Public	Surabaya	Girl	B03S
28	Year 12	MoEC-RT	Public	Surabaya	Boy	B04S
29	Year 12	MoEC-RT	Public	Surabaya	Boy	B05S
30	Year 12	MoEC-RT	Public	Surabaya	Boy	B06S
31	Year 12	MoEC-RT	Private	Surabaya	Girl	B07S
32	Year 12	MoEC-RT	Private	Surabaya	Girl	B08S
33	Year 12	MoEC-RT	Private	Surabaya	Girl	B09S
34	Year 12	MoEC-RT	Private	Surabaya	Boy	B10S
35	Year 12	MoEC-RT	Private	Surabaya	Boy	B11S
36	Year 12	MoEC-RT	Private	Surabaya	Boy	B12S
37	Year 12	MoRA	Private	Surabaya	Girl	B13S
38	Year 12	MoRA	Private	Surabaya	Girl	B14S
39	Year 12	MoRA	Private	Surabaya	Girl	B15S
40	Year 12	MoRA	Private	Surabaya	Boy	B16S
41	Year 12	MoRA	Private	Surabaya	Boy	B17S
42	Year 12	MoRA	Private	Surabaya	Boy	B18S
43	Year 12	MoRA	Public	Surabaya	Girl	B19S
44	Year 12	MoRA	Public	Surabaya	Girl	B20S
45	Year 12	MoRA	Public	Surabaya	Girl	B21S
46	Year 12	MoRA	Public	Surabaya	Boy	B22S
47	Year 12	MoRA	Public	Surabaya	Boy	B23S
48	Year 12	MoRA	Public	Surabaya	Boy	B24S

Table I. 2 *Four-Character Code for Junior High School Students*

No	Year	School type	School status	City	Gender	Student
1	Year 9	MoEC-RT	Public	Jombang	Girl	A01J
2	Year 9	MoEC-RT	Public	Jombang	Girl	A02J
3	Year 9	MoEC-RT	Public	Jombang	Girl	A03J
4	Year 9	MoEC-RT	Public	Jombang	Boy	A04J
5	Year 9	MoEC-RT	Public	Jombang	Boy	A05J
6	Year 9	MoEC-RT	Public	Jombang	Boy	A06J
7	Year 9	MoRA	Private	Jombang	Girl	A07J
8	Year 9	MoRA	Private	Jombang	Girl	A08J
9	Year 9	MoRA	Private	Jombang	Girl	A09J
10	Year 9	MoRA	Private	Jombang	Boy	A10J
11	Year 9	MoRA	Private	Jombang	Boy	A11J
12	Year 9	MoRA	Private	Jombang	Boy	A12J
13	Year 9	MoRA	Public	Jombang	Girl	A13J
14	Year 9	MoRA	Public	Jombang	Girl	A14J
15	Year 9	MoRA	Public	Jombang	Girl	A15J
16	Year 9	MoRA	Public	Jombang	Boy	A16J
17	Year 9	MoRA	Public	Jombang	Boy	A17J
18	Year 9	MoRA	Public	Jombang	Boy	A18J
19	Year 9	MoEC-RT	Private	Jombang	Girl	A19J
20	Year 9	MoEC-RT	Private	Jombang	Girl	A20J
21	Year 9	MoEC-RT	Private	Jombang	Girl	A21J
22	Year 9	MoEC-RT	Private	Jombang	Boy	A22J
23	Year 9	MoEC-RT	Private	Jombang	Boy	A23J
24	Year 9	MoEC-RT	Private	Jombang	Boy	A24J
25	Year 9	MoEC-RT	Private	Surabaya	Girl	A01S
26	Year 9	MoEC-RT	Private	Surabaya	Girl	A02S
27	Year 9	MoEC-RT	Private	Surabaya	Girl	A03S
28	Year 9	MoEC-RT	Private	Surabaya	Boy	A04S
29	Year 9	MoEC-RT	Private	Surabaya	Boy	A05S
30	Year 9	MoEC-RT	Private	Surabaya	Boy	A06S
31	Year 9	MoRA	Private	Surabaya	Boy	A07S
32	Year 9	MoRA	Private	Surabaya	Boy	A08S
33	Year 9	MoRA	Private	Surabaya	Boy	A09S
34	Year 9	MoRA	Private	Surabaya	Girl	A10S
35	Year 9	MoRA	Private	Surabaya	Girl	A11S
36	Year 9	MoRA	Private	Surabaya	Girl	A12S
37	Year 9	MoRA	Public	Surabaya	Girl	A13S
38	Year 9	MoRA	Public	Surabaya	Girl	A14S
39	Year 9	MoRA	Public	Surabaya	Girl	A15S
40	Year 9	MoRA	Public	Surabaya	Boy	A16S
41	Year 9	MoRA	Public	Surabaya	Boy	A17S
42	Year 9	MoRA	Public	Surabaya	Boy	A18S
43	Year 9	MoEC-RT	Public	Surabaya	Girl	A19S
44	Year 9	MoEC-RT	Public	Surabaya	Girl	A20S
45	Year 9	MoEC-RT	Public	Surabaya	Girl	A21S
46	Year 9	MoEC-RT	Public	Surabaya	Boy	A22S
47	Year 9	MoEC-RT	Public	Surabaya	Boy	A23S
48	Year 9	MoEC-RT	Public	Surabaya	Boy	A24S

Appendix J. Approval from Human Research Ethics Committee



Badrun.Kurnia

From:	donotreply@infonetica.net
Sent:	Thursday, 9 May 2019 4:07 PM
To:	Badrun.Kurnia@canberra.edu.au; Thomas.Lowrie@canberra.edu.au; Sitti.Patahuddin@canberra.edu.au; anisa.fatwasari@gmail.com; hafiyusholeh@gmail.com; Badrun.Kurnia
Cc:	humanethicscommittee@canberra.edu.au
Subject:	1576: Approved

Dear Badrun

The Human Research Ethics Committee has considered your application to conduct research with human subjects for the project "1576 - Indonesian High School Students' Statistical Literacy".

The Committee made the following evaluation: **Approved**

The approval is valid until: 31/12/2019

The following general conditions apply to your approval. These requirements are determined by University policy and the ***National Statement on Ethical Conduct in Human Research*** (National Health and Medical Research Council, 2007).

Monitoring

You must assist the Committee to monitor the conduct of approved research by completing project review forms, and in the case of extended research, at least annually during the approval period.

Reporting Adverse Events

You must report any unexpected adverse events or complications that occur anytime during the conduct of the research study or during the follow up period after the research. Please refer these matters promptly to the HREC. Failure to do so may result in the withdrawal of the Ethics approval.

Discontinuation of Research

You must inform the Committee, giving reasons, if the research is not conducted or is discontinued before the expected date of completion.

Extension of Approval

If your project will not be complete by the expiry date stated above, you must apply for extension of approval. This must be done before current approval expires.

Retention and Storage of Data

University policy states that all research data must be stored securely, on University premises, for a minimum of five years. You must ensure that all records are transferred to the University when the project is complete.

Contact Details and Notification of Changes

All email contact should use the UC email address. You should advise the Committee of any change of address during or soon after the approval period including, if appropriate, email address(es).

Please do not hesitate to contact us via email humanethicscommittee@canberra.edu.au if you require any further information.

All the best,

Hendryk Flaegel

Research Ethics & Integrity

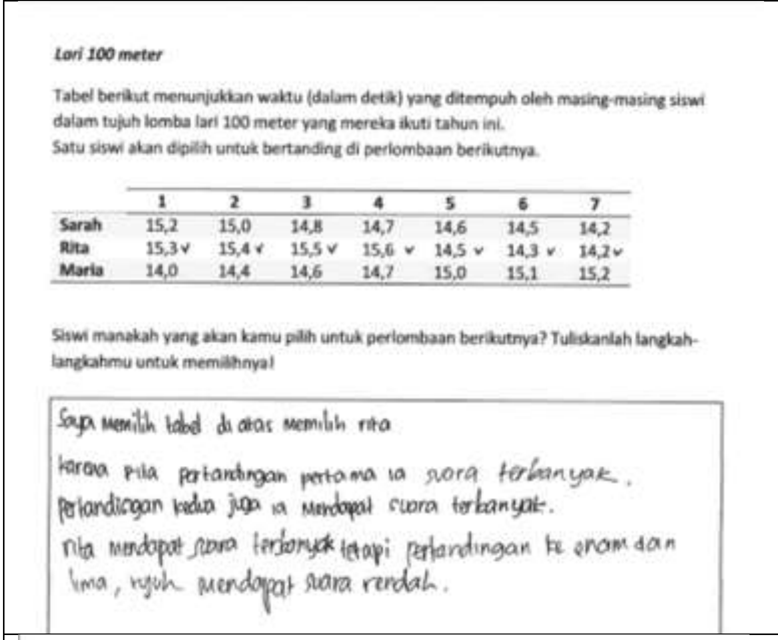
Research Services

University of Canberra

9 May 2019

Appendix K. Sample of Students' Responses Across Six Hierarchical Levels

Table K. *Example of Students' Works for The 100-Metre Race Item across Six Hierarchical Levels*

<table>
<tr><td>Students' works</td><td>Level and the description of student's work</td></tr>
<tr><td>


Lari 100 meter

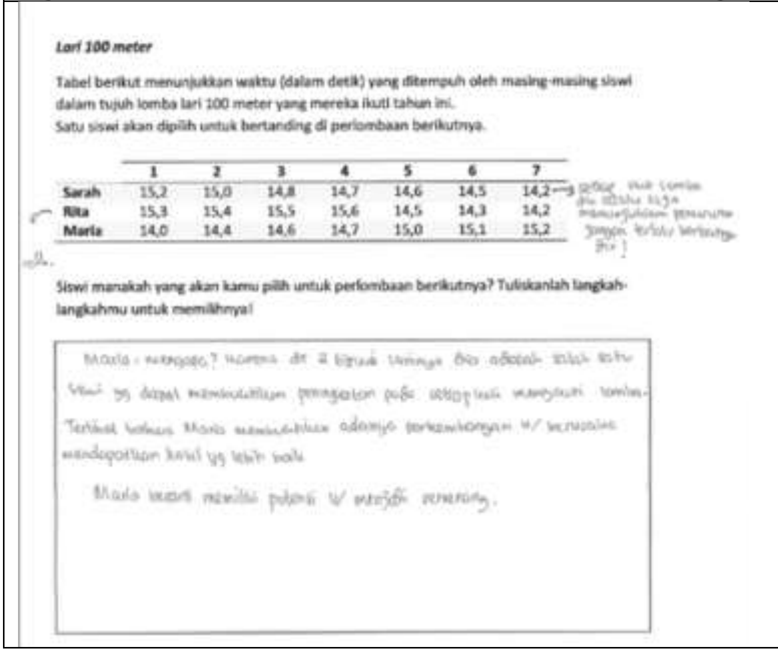
Tabel berikut menunjukkan waktu (dalam detik) yang ditempuh oleh masing-masing siswi dalam tujuh lomba lari 100 meter yang mereka ikuti tahun ini.
Satu siswi akan dipilih untuk bertanding di perlombaan berikutnya.

<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr>
<tr><td>Sarah</td><td>15,2</td><td>15,0</td><td>14,8</td><td>14,7</td><td>14,6</td><td>14,5</td><td>14,2</td></tr>
<tr><td>Rita</td><td>15,3 ✓</td><td>15,4 ✓</td><td>15,5 ✓</td><td>15,6 ✓</td><td>14,5 ✓</td><td>14,3 ✓</td><td>14,2 ✓</td></tr>
<tr><td>Maria</td><td>14,0</td><td>14,4</td><td>14,6</td><td>14,7</td><td>15,0</td><td>15,1</td><td>15,2</td></tr>
</table>

Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!

Saya memilih tabel di atas memilih rita
karena rila pertandingan pertama ia suora terbanyak.
pertandigan kedua juga ia mendapat suora terbanyak.
rita mendapat suara terbanyak tetapi pertandingan ke enam dan lima, tujuh mendapat suara rendah.
</td><td>

Level 1 (idiosyncratic)

This student chose Rita to compete in the upcoming championship as Rita got more votes in the first and second race.
</td></tr>
<tr><td>


Lari 100 meter

Tabel berikut menunjukkan waktu (dalam detik) yang ditempuh oleh masing-masing siswi dalam tujuh lomba lari 100 meter yang mereka ikuti tahun ini.
Satu siswi akan dipilih untuk bertanding di perlombaan berikutnya.

<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr>
<tr><td>Sarah</td><td>15,2</td><td>15,0</td><td>14,8</td><td>14,7</td><td>14,6</td><td>14,5</td><td>14,2</td></tr>
<tr><td>Rita</td><td>15,3</td><td>15,4</td><td>15,5</td><td>15,6</td><td>14,5</td><td>14,3</td><td>14,2</td></tr>
<tr><td>Maria</td><td>14,0</td><td>14,4</td><td>14,6</td><td>14,7</td><td>15,0</td><td>15,1</td><td>15,2</td></tr>
</table>

[illegible]

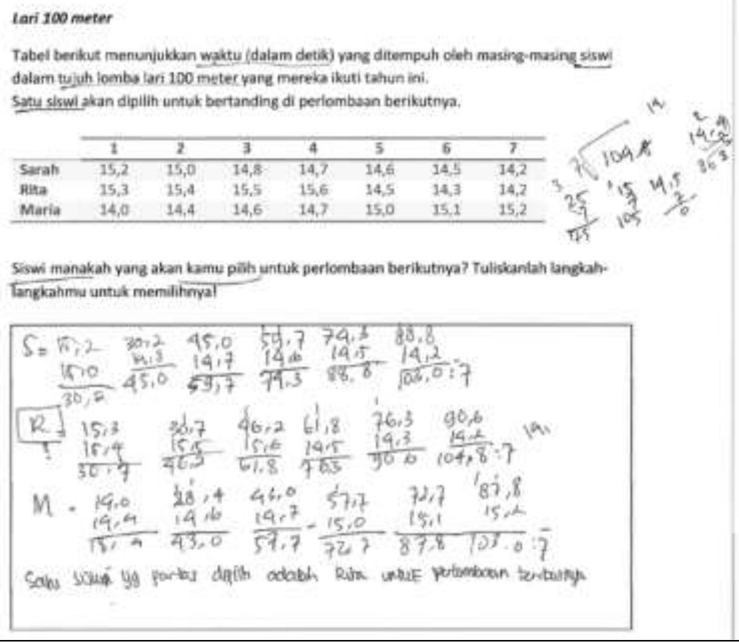
Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!

Maria : mengapa? karena dr [illegible] lombanya dia adalah salah satu siswi yg dapat membuktikan peningkatan pada setiap kali mengikuti lomba. Terlihat bahwa Maria menunjukkan adanya perkembangan u/ semakin mendapatkan hasil yg lebih baik

Maria akan memiliki potensi u/ menjadi pemenang.
</td><td>

Level 2 (informal)

This student chose Maria to compete in the upcoming championship. The method to select was by looking at Maria's trend which showed increasing in time across seven races meaning there was an improvement and thus Maria has potential to be a winner.
</td></tr>
</table>

<table>
<tr><th>Students’ works</th><th>Level and the description of student’s work</th></tr>
<tr><td>


Lari 100 meter

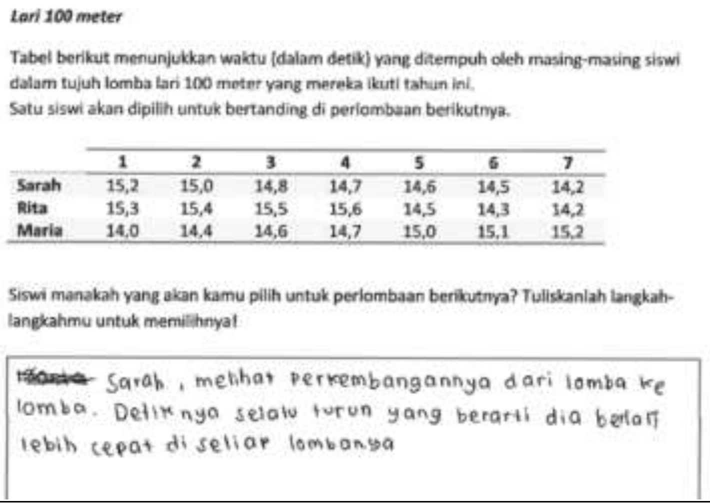
Tabel berikut menunjukkan waktu (dalam detik) yang ditempuh oleh masing-masing siswi dalam tujuh lomba lari 100 meter yang mereka ikuti tahun ini.
Satu siswi akan dipilih untuk bertanding di perlombaan berikutnya.

<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr>
<tr><td>Sarah</td><td>15,2</td><td>15,0</td><td>14,8</td><td>14,7</td><td>14,6</td><td>14,5</td><td>14,2</td></tr>
<tr><td>Rita</td><td>15,3</td><td>15,4</td><td>15,5</td><td>15,6</td><td>14,5</td><td>14,3</td><td>14,2</td></tr>
<tr><td>Maria</td><td>14,0</td><td>14,4</td><td>14,6</td><td>14,7</td><td>15,0</td><td>15,1</td><td>15,2</td></tr>
</table>

Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!

S = 15,2 + 15,0 = 30,2; 30,2 + 14,8 = 45,0; 45,0 + 14,7 = 59,7; 59,7 + 14,6 = 74,3; 74,3 + 14,5 = 88,8; 88,8 + 14,2 = 103,0 : 7
R. 15,3 + 15,4 = 30,7; 30,7 + 15,5 = 46,2; 46,2 + 15,6 = 61,8; 61,8 + 14,5 = 76,3; 76,3 + 14,3 = 90,6; 90,6 + 14,2 = 104,8 : 7
M = 14,0 + 14,4 = 28,4; 28,4 + 14,6 = 43,0; 43,0 + 14,7 = 57,7; 57,7 + 15,0 = 72,7; 72,7 + 15,1 = 87,8; 87,8 + 15,2 = 103,0 : 7
Satu siswa yg perlu dipilih adalah Rita untuk perlombaan berikutnya
</td><td>

Level 3 (inconsistent)

This student chose Rita to compete in the upcoming championship. The method to select was by finding the *mean* of time each runner needed to finish the race. The student found that Rita has the biggest mean compared to the other two runners.
</td></tr>
<tr><td>


Lari 100 meter

Tabel berikut menunjukkan waktu (dalam detik) yang ditempuh oleh masing-masing siswi dalam tujuh lomba lari 100 meter yang mereka ikuti tahun ini.
Satu siswi akan dipilih untuk bertanding di perlombaan berikutnya.

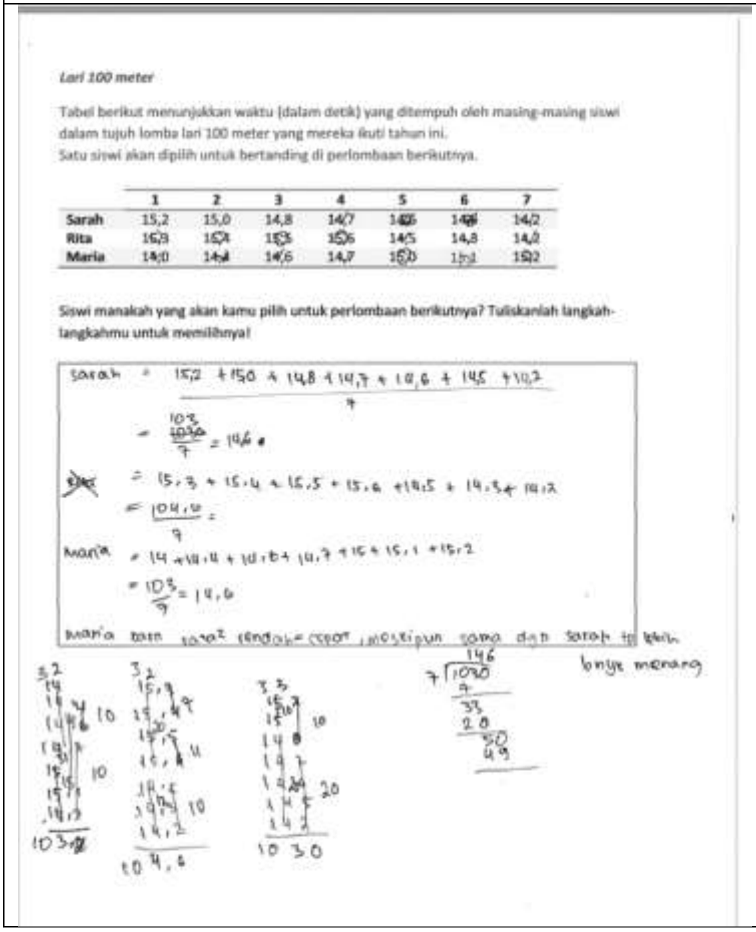
<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr>
<tr><td>Sarah</td><td>15,2</td><td>15,0</td><td>14,8</td><td>14,7</td><td>14,6</td><td>14,5</td><td>14,2</td></tr>
<tr><td>Rita</td><td>15,3</td><td>15,4</td><td>15,5</td><td>15,6</td><td>14,5</td><td>14,3</td><td>14,2</td></tr>
<tr><td>Maria</td><td>14,0</td><td>14,4</td><td>14,6</td><td>14,7</td><td>15,0</td><td>15,1</td><td>15,2</td></tr>
</table>

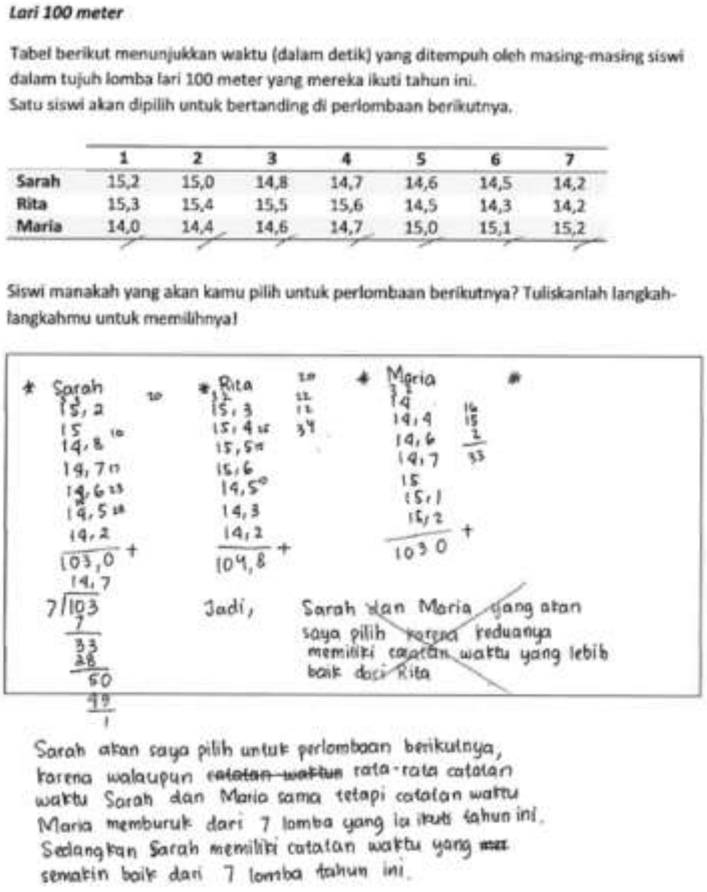
Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!

~~Maria~~ Sarah, melihat perkembangannya dari lomba ke lomba. Detiknya selalu turun yang berarti dia berlari lebih cepat di setiap lombanya
</td><td>

Level 4 (consistent non-critical)

This student chose Sarah to compete in the upcoming competition. The selection method applied was looking at Sarah’s trend which showed decreasing in time across seven races meaning getting quicker.
</td></tr>
</table>

Students’ works	Level and the description of student’s work
 *Lari 100 meter* Tabel berikut menunjukkan waktu (dalam detik) yang ditempuh oleh masing-masing siswi dalam tujuh lomba lari 100 meter yang mereka ikuti tahun ini. Satu siswi akan dipilih untuk bertanding di perlombaan berikutnya. 1 2 3 4 5 6 7 Sarah 15,2 15,0 14,8 14,7 14,6 14,5 14,2 Rita 15,3 15,4 15,5 15,6 14,5 14,8 14,2 Maria 14,0 14,4 14,6 14,7 15,0 15,1 15,2 Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya! sarah = 15,2 + 15,0 + 14,8 + 14,7 + 14,6 + 14,5 + 14,2 / 7 = 103 ~~1030~~ / 7 = 14,6 ~~Rita~~ = 15,3 + 15,4 + 15,5 + 15,6 + 14,5 + 14,3 + 14,2 = 104,4 / 7 = Maria = 14 + 14,4 + 14,6 + 14,7 + 15 + 15,1 + 15,2 = 103 / 7 = 14,6 maria dan sarah2 rendah= cepat , meskipun sama dgn sarah tp lebih bnyk menang 146 7)1030 7 33 28 50 49 103,0 104,6 1030	*Level 5 (critical)* This student chose Maria to compete in the upcoming competition. The method applied was finding the *mean* of time each runner needed to finish the race. Realizing that there were two runners (Sarah and Maria) having the same *mean* of time, the student compared the mode for Sarah and Maria. The student found that Maria won the races more than Sarah.

<table>
<tr><th>Students’ works</th><th>Level and the description of student’s work</th></tr>
<tr><td>


Lari 100 meter

Tabel berikut menunjukkan waktu (dalam detik) yang ditempuh oleh masing-masing siswi dalam tujuh lomba lari 100 meter yang mereka ikuti tahun ini.
Satu siswi akan dipilih untuk bertanding di perlombaan berikutnya.

<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr>
<tr><td>Sarah</td><td>15,2</td><td>15,0</td><td>14,8</td><td>14,7</td><td>14,6</td><td>14,5</td><td>14,2</td></tr>
<tr><td>Rita</td><td>15,3</td><td>15,4</td><td>15,5</td><td>15,6</td><td>14,5</td><td>14,3</td><td>14,2</td></tr>
<tr><td>Maria</td><td>14,0</td><td>14,4</td><td>14,6</td><td>14,7</td><td>15,0</td><td>15,1</td><td>15,2</td></tr>
</table>

Siswi manakah yang akan kamu pilih untuk perlombaan berikutnya? Tuliskanlah langkah-langkahmu untuk memilihnya!

* Sarah
15,2
15
14,8
14,7
14,6
14,5
14,2 +
103,0
14,7
7/103
7
33
28
50
49
1

* Rita
15,3
15,4
15,5
15,6
14,5
14,3
14,2 +
104,8

* Maria
14
14,4
14,6
14,7
15
15,1
15,2 +
1030

Jadi, ~~Sarah dan Maria yang akan saya pilih karena keduanya memiliki catatan waktu yang lebih baik dari Rita~~

Sarah akan saya pilih untuk perlombaan berikutnya, karena walaupun ~~catatan waktu~~ rata-rata catatan waktu Sarah dan Maria sama tetapi catatan waktu Maria memburuk dari 7 lomba yang ia ikuti tahun ini. Sedangkan Sarah memiliki catatan waktu yang semakin baik dari 7 lomba tahun ini.
</td><td>
Level 6 (critical mathematical)

This student chose Sarah to participate in the upcoming competition. The selection method was finding the *average (i.e., mean)* of time each runner needed to finish the race. The student compared the trend for Sarah and Maria after realizing that two runners (Sarah and Maria) had the same *mean* of time. The student discovered that Sarah’s trend throughout seven races is getting quicker whereas Maria is slowing down.
</td></tr>
</table>

Appendix L. Distribution of Students' Levels in All Components for Communicating Items

Table L. *Distribution of Students' Levels in All Components for Communicating Item across the Hierarchy by Percentage*

		Lower group			Upper group		
Item (Year)	Compo-nent	*L1*	*L2*	*L3*	*L4*	*L5*	*L6*
YouTube viewers (Year 9)	TnC	0%	4%	23%	60%	13%	0%
	Rep	0%	4%	21%	65%	10%	0%
	SnM	0%	4%	17%	60%	19%	0%
YouTube viewers (Year 12)	TnC	0%	0%	6%	63%	29%	2%
	Rep	0%	0%	0%	67%	31%	2%
	SnM	0%	0%	4%	50%	44%	2%
Domestic Waste (Year 9)	TnC	2%	6%	13%	48%	31%	0%
	Rep	2%	4%	13%	68%	13%	0%
	SnM	0%	6%	19%	63%	10%	2%
Domestic Waste (Year 12)	TnC	0%	0%	13%	50%	35%	2%
	Rep	0%	0%	10%	69%	19%	2%
	SnM	0%	0%	2%	67%	27%	4%