

Uncovering students' views through cognitive interviews: Validation procedure for statistical literacy test items

Achmad Badrun Kurnia^{1*}

¹University of PGRI Jombang

Abstract

High school students should be prepared and reasonably expected to be critical data-based information readers. Their critical responses are a strong predictor of statistical literacy (SL) and thus a cohesive picture of it is needed to better inform educational stakeholders. In response to this notion, it is essential to use quality instruments to measure students' SL. This study aims to describe the careful and systematic processes of developing and validating an alternative assessment to measure high school students' SL. Each item in the assessment assessed one of the SL skills (interpreting, communicating, evaluating, and decision-making) and the item features contain the three SL components. The evidence for reliability and validity was collected through a series of cognitive interviews (cognitive interviews I and II). Cognitive interview I involved two 9-grade and two 12-grade students with average levels of knowledge in each grade. In comparison, cognitive interview II involved six 9-grade and six 12-grade students with different levels of knowledge (2 low, 2 medium, and 2 high level students). Furthermore, cognitive interview II was administered after the students solved the test items under test condition. Results confirm that ten items of the SL instrument were valid to be used for data collection. Implication of this study includes a comprehensive process of validating SL test items that contribute to the assessment theory, particularly on SL assessment in Indonesia.

Keywords: Statistical Literacy, Test Items, Validation, Cognitive Interview

Most high school students can actively access media information in today's data-driven environment, which frequently challenges them to respond to data-based arguments, opinions, and information. Their response to this information is dependent upon their comprehension of the statistical concepts and context. They are further expected to possess critical response to such information, making them knowledgeable consumers. However, most high school students might find it challenging to critically assess and comprehend information derived from data, which could cause them to be misinformed (Budgett & Pfannkuch, 2010). Indonesian high school students (aged 15-16 years old), for example, performed poorly on data-based problems over decades (OECD, 2004; 2014; 2023). As a result, SL is obviously essential for Indonesian students, and as such, an assessment that could further examine their SL needs to be carefully designed.

Despite its importance, the definition of SL has been in a debate among experts, yet encompasses wide-ranging and complex skills beyond procedural calculations (Bailey & McCulloch, 2023; Budgett & Pfannkuch, 2010; Budgett & Renelle, 2023; Budgett & Rose, 2017; Büscher, 2022; Gal, 2019; Helenius et al., 2020; Jureckova & Csachova, 2020; Sabbag et al., 2018; Schield, 2017; Sharma, 2017). First and foremost, Wallman (1993) defined SL as the capacity to comprehend and assess statistical findings that are present in our everyday lives for professional and personal decision-making. Additionally, Gal (2002) proposed the definition of SL as the ability to communicate reactions and opinions orally or in writing, as well as to comprehend statistical information and data-related arguments. Nevertheless, according to

* Corresponding author:
Email Address: badrun.bbest@yahoo.com

Ben-Zvi and Garfield (2004), SL is the capacity to comprehend the meaning of basic statistical terminology, analyze various data representations, and comprehend the application of basic statistical symbols.

Considering the broad meaning of SL, the assessment development begins with defining the SL construct to be assessed. The process of defining SL began with classifying the existing definition into two perspectives: data producers and data consumers. As more high school students become statistical consumers, advanced classification was then applied to the data consumers' perspective. However, the SL definition from this perspective was extensive and needed to be carefully synthesized. The synthesis process focused on a wide range of data consumption skills required by data consumers, including students. From this perspective, statistically literate students must demonstrate skills such as drawing qualitative meaning from data often presented quantitatively (Cui et al., 2023; Wallman, 1993), effectively presenting or discussing statistical information with others by selecting the most significant data (Cui et al., 2023; Gal, 2002; Sutherland et al., 2022), arguing for statistical claims or arguments with fair and critical data-based proof (Budgett & Rose, 2017; Büscher, 2022; Guler et al., 2016; Koga, 2022), and making informed decisions with relevant arguments (Budgett & Renelle, 2023; Franklin et al., 2005). Overall, four critical skills for statistically literate data consumers have been identified: **interpreting**, **evaluating**, **communicating** and **decision-making**. Figure 1 presents the working SL definition for this study, particularly on those four skills.

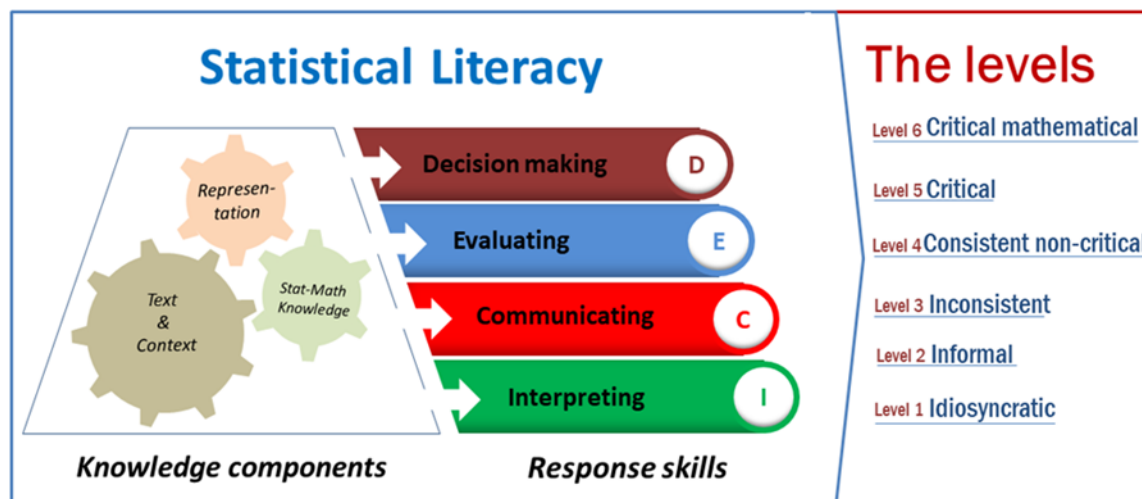


Figure 1. SL Assessment Framework (Kurnia, et al., 2024)

The SL framework presented in Figure 1 then served as the foundation to develop a quality instrument to assess students' SL. One of many methods, cognitive interviews could be used to both improve the quality of the test items and anticipate students' response. In other words, conducting cognitive interviews with students will contribute to the validity of both the assessment items and the conjectured students' thought processes (Desimone & Le Floch, 2004). Furthermore, Tiruneh et al. (2017) outline some benefits of instrument piloting using cognitive interviews. First, it demonstrates if the students understand the problems. Second, it provides evidence of how students respond to the items. A cognitive interview gathers more verbal data regarding how the students responded to the test items (Beatty & Willis, 2007; Conrad & Blair, 1996; Desimone & Le Floch, 2004). Third, it seeks to determine whether the test items can facilitate various responses from students. It is then evident that the students'

various responses can be anticipated (Sabbag et al., 2018) and inaccurate responses brought on by instrument errors can be reduced (Ziegler & Garfield, 2018).

Theoretically, cognitive interviews can be conducted through four problem-solving processes for obtaining students' responses. Those processes are 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; Shafer & Lohse, 2005; Willis, 1999, 2005). These processes could help ensure that the items measure the desired SL skills and look into students' general comprehension and misconceptions when answering SL items. Their misconceptions are sources for item refinement.

This study was conducted as part of a larger PhD research which investigated the SL of Indonesian 9th and 12th grade students. The research question being addressed in this study is "What issues with item clarity do cognitive interviews reveal with regard to text and context, representation, and questions?" Given this research question, the purpose of this study is to provide evidence of the clarity, relevance, comprehensibility and cognitive demands of the items to assess the intended SL skill. In other words, this study uncovers students' views (such as interpretation and cognitive load) through cognitive interviews (how students interpret, process, and respond to each item) to enhance validity and usability.

METHODS

This rigorous process of assessment development to provide reliability and validity evidence involved systematic item piloting and refinements through two pilot interviews. Before piloting, 14 statistical items were theoretically developed, modified (Figure 2 provides an instance of the item), and their content validity was checked by three experts. There was an agreement among those three experts on the skill assessed by each item and the items' contents were considered appropriate for both grades 9 and 12. Among those 14 items, four items assess interpreting skill, two items assess communicating skill, five items assess evaluating skill, and three items assess decision-making skill. Using these items, possible insights from the students' viewpoints were examined.

The 100-metre race

The following table gives the times (in seconds) that each girl has recorded for 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? Show your strategies!

Figure 2. The Initial Version of *The 100-metre Race* Item

Participants

The participants in this study were students in grades 9 and 12. Given that two cognitive interviews would be conducted, each interview had different participants. The participants for cognitive interview I were four students (half of them were grade 9 students), while the participants for cognitive interview II were 12 students (half of them were grade 9 students). The involvement of 16 students in total was considered common and enough for identifying response patterns and misunderstanding trends among students. Furthermore, the involvement of more students in the cognitive interview II was expected to show that there were no more new issues, meaning the students' responses are saturated.

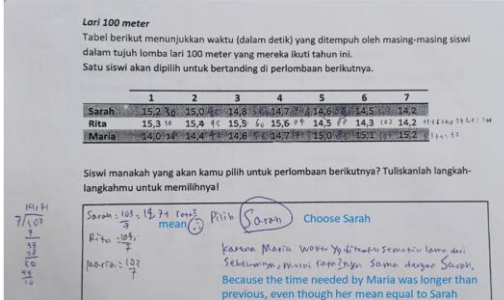
With the support of their mathematics teachers, two 9-grade and two 12-grade students with an average level of knowledge voluntarily participated in *cognitive interview I*. Although these students were selected by the teachers, there were some steps to minimize bias and to enhance the credibility of findings. First, the average students became the inclusion criteria and thus those who belonged to high and low achievers were omitted. Second, those students were selected to illustrate the typical thought process and responses of 9 and 12 grade students with average levels of knowledge. Third, those students were selected from different grades and cohorts to increase diversity and consistency and reduce the contextual bias and uncontrolled variability. The empirical insight obtained from the students' responses was expected to further enrich the items that had previously been theoretically developed.

Following *cognitive interview I*, *cognitive interview II* was conducted with 12 students (six 9-grade and six 12-grade students). *Cognitive interviews I* and *II* differed in that the latter was conducted after students had completed the 14 items under test conditions. The purpose of *cognitive interview II* was to clarify students' written responses. These students agreed to voluntarily participate in the interview. Further, those 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 levels of knowledge. The wider the range of responses emerging from students, the better the researcher would predict the spectrum of students' responses in the actual test.

Research Instruments

Before conducting interviews, a cognitive interview protocol was developed, and it consisted of the interview technique and the interview script. The interview technique described the preparation and interview setting, while the interview script included questions regarding four problem-solving processes as described in the previous section. In order to cover these four problem-solving processes, the interview questions were developed along with appraisal statements that addressed the students' responses. For instance, "*please kindly read the question out loud!*" was the first instruction asked to start the interview and investigate students' item comprehension; "*when you start answering, what are you thinking?*" was a probing question aimed at determining the students' stages in solving the problem; and "*your reaction is helpful, thank you*" was an encouraging reaction to students' answers. Table 1 presents script examples for cognitive interviews.

Table 1. Script Examples for Cognitive Interviews I and II

Step	Script for Cognitive Interview I	Script for Cognitive Interview II
Beginning the interview (4 minutes)	Before we begin this interview, let's have a proper introduction. My name is [insert name], and you are [insert interviewee's name]. It's nice to meet you. First of all, thanks so 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. 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].	Before we begin this interview, let's have a proper introduction. My name is [insert name], and you are [insert interviewee's name]. It's nice to meet you. 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, [insert interviewee's name]? Thank you. During this interview, please think aloud as you were solving the problems. That means say anything—whatever you think; I'm interested in hearing all of 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 kindly read the question out loud! 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.	 Comprehension of the item and retrieval of relevant information: Can you tell me what you can understand from The 100-metre Race item? Please say it aloud! 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. Judgement-making based upon the recall of knowledge: Next, what is this question asking? Probe: Okay, can you explain the way you chose the best runner? That's great.

Step	Script for Cognitive Interview I	Script for Cognitive Interview II
	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)	The process of mapping the answer onto the reporting system: 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 concludes this interview, and I will now stop the recording at time [insert time].	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].

The probing questions in the scripts for *cognitive interviews I* and *II* differed slightly. In *cognitive interview I*, the interviewer simply followed the students' thought processes as they interacted with each item, and thus the probing was not overly intense. On the other hand, the purpose of the *cognitive interview II* probing was to elucidate the students' thought processes during the written exam. Their written answers were analyzed before the interview, and conjectures were made about their thought process. There was variation in the probing questions because each student's thought process and written response was unique. By concentrating only on each student's potential misunderstandings regarding the three SL components and SL skill, the interviewer attempted to preserve consistency.

Data Collection

In this study, two cognitive interviews were used to investigate students' understandings of each item's features developed under the SL framework (see Figure 1). The item's features include 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. The two cognitive interviews are described below:

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 9-grade 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 this interview 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.

The procedures for *cognitive interview II* were the same as those for *cognitive interview I*. The interviews were conducted across four days. Each of the 12 students was interviewed, except for one 9-grade 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. Technically, this interview was conducted in a 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 processes eventually improved the depth and richness of the information gathered. This information proved crucial in helping to confirm and evaluate the previous conjectures about students' cognitive processes.

Data Analysis

After finishing each cognitive interview, the researcher began searching for possible clues for test items refinement by listening to the audio recording. The researcher conducted this work with the help of a research assistant. Both the research assistant and the researcher conducted thematic analysis on the students' responses to identify possible issues in the wording in the text, the graphical and tabular displays and the questions that were worthwhile for item refinement. They conducted the analysis together from each cognitive interview's recording. This process was administered to avoid errors in measurement instruments that could result in low reliability (Weathington et al., 2010). This process of piloting the test items (*cognitive interview I* and *II*), analyzing students' responses, refining the items, and finalizing the instruments contributed to instrument validity.

RESULTS AND DISCUSSION

The results from analyzing the data from both cognitive interviews are presented separately, followed by the discussion of some key findings.

Cognitive Interview I Result

The intention to conduct this *cognitive interview I* was to look for some potential insights coming from the eyes of students regarding their interpretation of the items including texts informing the context of the problems, the graph or table displaying the context-related data, and the question assessing one of the skills. Understanding the problem from the students' eyes could minimize the incorrect answers caused by the instrument errors. Table 2 shows the various responses from interviewing four students about *The 100-metre Race* item (assessing decision-making skill), while explanations on how these responses motivated the refinement follow.

Table 2. Cognitive Interview I Result for *The 100-metre Race* Item

Feature	Students' Comprehension	
	Grade 9	Grade 12
Text and Context	Students read the texts and directly relate the information to the table	Students read the texts carefully and restate to express understanding
Table	Table reading ▪ One of them in this table must be the best	Table reading ▪ The numbers in the tables show the times

Feature	Students' Comprehension	
	Grade 9	Grade 12
	- 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	- Sarah in the 1st race is 15.2 secs, Rita 15.3 secs, and Maria 14.0 secs and Maria is the quickest because only 14.0 seconds - What do these numbers [pointing 1 to 7] mean? - What do these numbers mean? 14,2?
Question	The fastest or the most stable runner - I selected Maria, oh, Rita because she was the most stable from 1st to 5th race The biggest time - The winner in the 7th race is absolutely Maria and in the 2nd race is Rita The fastest mean - So, finding 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 select the quickest - The mean is finding the total time and divided by how many races

After analyzing the students' interviews for all 14 items, three types of refinements were made. The first type was applied within the contextual texts. These changes varied from merely adding a comma (punctuation), replacing-adding the wording or numbers to creating contextual sentences to provide students with a clearer context. The second type of amendment relates to the items' graphical or tabular displays, such as applying small changes to the one column caption, adjustment to one bar in the graph and the x-axis. Finally, the third type of items' refinement was applied to the question wording. This edit was applied to seven items varying from changing a single word, adding some words, omitting words, and question modification. The refined items were then used for *cognitive interview II*.

Cognitive Interview II Result

Of the twelve students, the interview results of three students were analyzed to show their different levels of understanding: a solid understanding and common misunderstandings. In addition, the interview confirmed whether final refinements were needed. Figure 3 exemplifies the written work of three students showing solid understanding and common misunderstandings on *The 100-metre Race* item. Amal demonstrated a solid understanding for this problem, applying two methods of selection. From his written answer, it was predicted that he started working by calculating the total time each runner had for seven races. The calculation method he applied indicates his great understanding of number sense, adding on only all the integers and followed by adding on the decimal numbers. Once he found the total time, he continued calculating the mean of time each runner had. Finding the fact that there were two runners having the same mean, Amal used the trend to further decide which runner should compete in the next championship. This way of thinking eventually led Amal to choose Sarah among the other two runners.

In comparison to Amal, Zainudin could find the total time and mean each runner had. However, he exposed common misunderstandings when dealing with the best runner. According to him, the best runner is the one whose mean is the biggest. Therefore, he selected Rita instead of Sarah or Maria as Rita has the biggest mean. Moving to Oktav's response as further comparison, he worked with mode as a means to select the best one among three runners to compete in the next championship. The checklists in the table indicated her thoughts or selection technique. Oktav chose the one with the highest number (in fact it is the times in seconds) as the winner in each race. This technique has led Oktav to choose Rita as she won (was check-listed) four out of seven races.

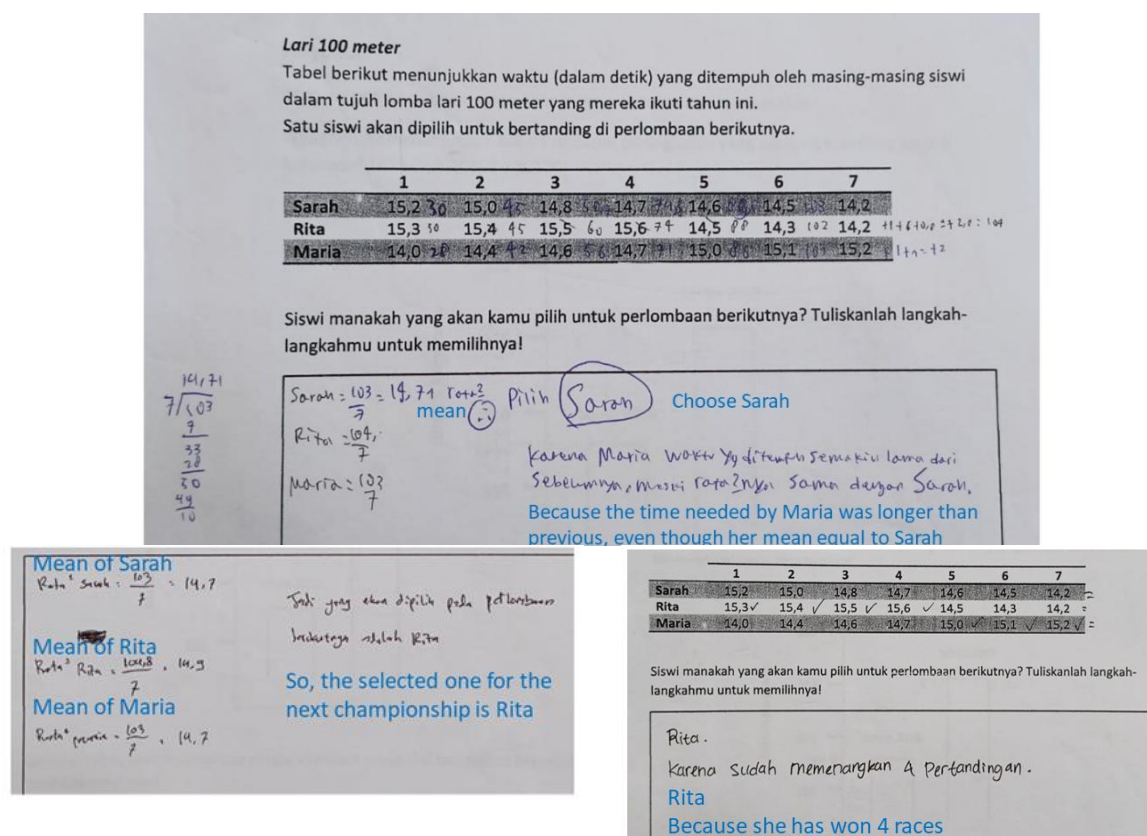


Figure 3. Three Samples of Students' Responses: Amal (above), Zainudin (below left), and Oktav (below right)

The process of a thorough analysis of students' written responses as above was applied to all 12 students' responses. Such a process has led the researcher to a hypothetical conjecture on what they did and did not understand. Afterwards, a unique cognitive interview protocol was developed to investigate students' thought process as deeply as possible while interacting with the fourteen items. From the interview, students' trajectory from item comprehension to answering process could be traced and identified. The results of interviewing those three students are presented.

Amal's Solid Understanding

Beginning with Amal, a 12-grade student, he showed a solid understanding in his written response. Amal was successful in making sense of the context of running competition and a series of data expressed in the table. His technique to select one best runner for the next championship revealed his expertise in number sense and his conceptual understanding of data measures of center and spread. It was predicted that he first tried selecting the best runner using *mean* and he then employed *trend* once he realized that Sarah and Rita had the same *mean*. The interview confirmed his solid understanding; however, his strategy was found another way around.

When asked to explain his process of item comprehension, Amal appeared to incorporate the information in the texts and data in the table. When he was asked to explain what he understood from the texts above the table, he replied "*I just read and there was no crucial information on the texts*". Although Amal (A) stated that to the interviewer (I), it did not indicate that he had no understanding of it. He then directly looked 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 the times] are the records for 100 meters, the finish [time is] 15.2 seconds, 15.3, and 14.0 [pointing Sarah, Rita, and Maria 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 at the beginning.
- I : Then please explain what you were doing in the beginning!
- A : Firstly, I observed the numbers, the finish times; from 1 to 7, it is a running race; so, it should be the fastest completion time. I observed that Sarah the more [race], the faster [her finishing time]. Rita was the same as Maria. Maria, the more [race], the slower [her finishing time]. After that, [I] made hypothesis that the answer is Sarah. Then, to prove the hypothesis, [I] calculated the total time for all students and mean for Sarah. Sarah and Maria were the same, and Rita's mean is longer. Although Maria and Sarah were same, Maria the more [race] the slower [her finishing time], so I chose Sarah.

His explanation revealed that he started selecting one best runner from observing the three runners' *trend* of finishing times over seven races and then making a hypothesis about who the best runner was. He hypothesized that Sarah was the best runner to select as she was getting faster through the seven races. To prove his hypothesis that Sarah was the best candidate for the championship, Amal calculated the mean for all the three runners for comparison. Finding the fact that Sarah's *mean* was both the smallest and similar to Maria, he then stepped back to compare the *trend* of both. Finally, Amal made his choice for Sarah as her getting-faster trend, while Maria was in the opposite with her getting-slower trend.

With respect to Amal's calculation strategy to calculate the *mean* for three runners, he performed an excellent *number sense*. When Amal was asked to explain his calculation, he answered "*in calculating, I did not like if the total is not an integer, I didn't consider the number after decimal point. It was calculated at the end like $0.4+0.6=1$, $0.7+0.1+0.2=1$, and Rita is the most difficult to count*". From his explanation, it can be interpreted in this way; for example, in the case of Maria, the calculation was first applied to the integers side by side. Only the integers in the first race (14) were added to the second race (14) and the result (28) was written in between. This 28 was further added to the time in the third race (14), and it resulted in 42 that was written between times in the second and third race. This addition continued to the last race and resulted in 101 for seven races. Not finished yet, it was the turn to add all the decimals. The decimals in the second (.4) and third race (.6) were added to make 1 and the decimals in the fourth (.7), sixth (.1), and seventh race (.2) were added to result in 1. Therefore, the total time for Maria is $101+1+1 = 103$ (see again Amal's written work in Figure 3). This number sense-based calculation was also applied to the other two runners.

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

Zainudin's Common Misunderstanding

Zainudin, a 9-grade student, showed contextual misunderstanding in interpreting the times three runners need to reach the finish line. Although his calculation for the *mean* each runner was correct, his contextual understanding failed him in selecting the best runner. During the interview, he said that he started comprehending the items from the table and the context and then continued reading the question. What Zainudin (Z) was actually thinking of was told to the interviewer (I).

- I : Where did you start comprehending the item?
 Z : From the table.
 I : What did information you get from the table?
 Z : There were three students competing in running. For seven times and got different results.

From his explanation, Zainudin seemed to have no problem understanding the context and properly interpreting the data in the table based on the context. He further showed his understanding of statistics and mathematical knowledge by using *mean* as a method of selection.

- Z : Adding on 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 is next?
 Z : Divided by seven because there were seven competitions.
 I : To find what?
 Z : To find the mean.

At this point, all his arguments were still making sense as he utilized *mean* as a method of selection. The total time for each runner was correct, the *mean* calculated was also correct, Sarah 14.7, Rita 14.9, and Maria 14.7. However, a common misunderstanding was just found when he started explaining the selection process.

- R : What do you think of after finding the mean for those three runners?
 Z : I directly thought that Rita should compete in the next championship because she has 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 clear that he actually had a misconception on the context of time three runners need to reach the finish line. He thought that the bigger the *mean*, the better the runner, not otherwise.

To this end, it could be concluded that Zainudin could understand from the texts and the table that this item was about three runners who completed seven races and their times to reach the finish line were recorded in the table. In addition, he also had a correct understanding that the question assesses his *decision-making* skill. Zainudin's choice to use *mean* as a selection method and to apply *the mean* formula did not show any signs of error. Unfortunately, he falsely interpreted the calculated *mean* he got for the three runners by choosing the biggest as the best. Thus, Zainudin's misunderstanding was purely caused by his lack of knowledge and not by the item's ambiguity.

Oktav's Common Misunderstanding

Oktav, a 9-grade student, started comprehending the item from the question first and then continued to the texts and the table. She understood the context of three runners of 100 meters for seven races. However, she could not recognize the numbers as the times each runner spent for 100 meters; instead, she recognized it as a score. She thus had a common misunderstanding and concluded that the winner in each race should have the highest score among others. She made a checklist for the winner in each race and eventually selected a runner who won the most races to compete in the next championship. Her misunderstanding was reflected through the interview below:

- I : From the table and the text, what can you understand?
 O : Here is Rita [pointing Rita's time] who is the biggest, from Maria
 I : If I ask you, what is it 1, 2, until 7?
 O : The competition 1, 2, and 3
 I : Now, please explain the checklists [you made] in the table?

- O : This is, here in the first competition, the winner is Rita, the second competition is also Rita, the third is Rita, the fourth Rita, and the fifth is Maria, the sixth is Maria, and the seventh is Maia.
- I : So, the checklists show the winner, how do you know that in the first competition the winner is Rita?
- O : I thought these [pointing the times] are scores
- I : Score? Ok it means you consider 15.3 as 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 idea of *mode* in statistics: whoever won the most races was selected to compete in the next championship. As she thought in that way, she identified the winner in each race and found that Rita won four times, Maria won three times, while Sarah never won at all. Without involving another reason, Oktav selected Rita as she won four times.

However, the interviewer did not want to end the interview at this point. There was curiosity as to whether the context was really difficult to understand or it was just Oktav's momentary lack of knowledge. For this purpose, the interviewer asked Oktav to read the texts again and explain what she could comprehend. As a result, she could correct her misunderstanding by herself that the numbers in the table are times instead of scores. By having this understanding, she corrected herself 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 comprehend it
- O : [Reading the text above the table]
- I : Now I ask again, what is it 1 to 7?
- O : 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 Sarah]
- O : Hmmm, wait a minute, it's wrong
- I : What do you mean? What do you understand now?
- O : This should be times, showing times in second. So, must be her [pointing Maria] because her time is the quickest, the smallest.
- I : Can you please explain again?
- O : In the text, it was written "this table show times (in seconds)" so these are times. First, I thought it was scores. So, the winner should be Maria because her time is the smallest

From this interview, it can be concluded that Oktav's misunderstanding was not caused by the item, but by herself. Therefore, no refinements were needed for this item.

Based on the findings of students' *cognitive interview II* on the item *The 100-metre Race*, it could be concluded that this item and its features (text, table, and question) are comprehensible for the students. Students' misunderstandings or incorrect responses were merely caused by their lack of knowledge such as their failures to make sense of the context of running competition and the table in the way that the winner should be those with the smallest time, total time, *mean*, or getting-faster trend. Therefore, no revision is needed, and this item is valid.

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 contextual texts provided. 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. Some items had additional instructions included encouraging students to give lengthy answers.

In summary, *cognitive interview II* resulted in 10 test items for the actual data collection. Four of the initial items were removed because the findings from both cognitive interviews proved that they could not be improved to cover various responses. The 10 selected items comprised four interpreting items, two communicating items, two evaluating items, and two decision-making items.

Discussions

This study has presented in great detail the stages of validating test items through cognitive interviews. The reliability and validity evidence, from a qualitative perspective, revolved within these stages. Further, the discussion continues on how the validation process can be improved for large-scale assessment and the importance of having accurate conjectured responses. Each of which is discussed in the following.

Adaptation for Large-Scale Assessment

The validation process employed in this study was considerably sufficient for a qualitative small-scale assessment but might not be for a large-scale assessment. Recalling that the items validated in this study were intended to cross-sectionally assess 9th and 12th grade students' SL, this study is a qualitative and small-scale assessment. Similar works on qualitative assessment development have been conducted, for example, by Mooney (2002) for Statistical Thinking, Sharma et al. (2013) for Statistical Literacy, and Ralston et al. (2018) for Algebra. Most of these studies employed a somewhat similar qualitative approach to develop and validate the items or frameworks. All those assessments were conducted with a few students whose responses were primarily analyzed qualitatively. However, it appears important to expand the current study and aim for a large-scale assessment such as throughout Indonesia. In this case, a quantitative approach might be required. Large-scale assessments such as Guidelines for Assessment and Instruction in Statistics Education (GAISE), Programme for International Student Assessment (PISA), and Trends in International Mathematics and Science Study (TIMSS) employed a combination of both qualitative and quantitative approaches in all stages from development to assessment conduct. Meanwhile, Watson and Callingham (2003) and Callingham and Watson (2017) applied a quantitative-dominated approach in all such stages. Thus, it is expected that if there is an intention to replicate or adapt the validation method as that in this study for a large-scale assessment, the combination of both qualitative and quantitative approaches appears very constructive.

The Importance of Accurate Conjectures

In addition to its purpose, this study provided insight into the actual data gathering process, including the conjecture of students' responses. In particular, this study revealed that cognitive interviews helped refine the initial conjectures regarding students' answers. Additionally, *cognitive interview II* expanded on the range of responses that students might provide. These various responses further provide a clue on how the students from different cohorts respond to test items and how these responses could be graded and categorized across hierarchical levels. This process aligned with previous assessments such as Mooney (2002), Watson & Callingham (2003), and possibly PISA and TIMSS.

Moreover, this study demonstrated that the conjectures were not only limited to students' correct or incorrect answers. Indeed, the conjectures could be related to students' cognitive processes, reflected through four problem-solving processes. The four problem-solving processes employed in the cognitive

interview follow Desimone & Le Floch (2004), Shafer & Lohse (2005), and Willis (1999, 2005), involving: students' attempt to comprehend each item, retrieve relevant information, recall related knowledge and map strategies to answer the questions. By having these conjectures, the test developer could more accurately predict the test administration, even before the piloting.

CONCLUSION

This study has described the details of piloting test items through cognitive interviews. The results of this study provided evidence on the importance of conducting a pilot study before data collection, particularly through cognitive interviews. The results revealed some issues with item clarity, including unclear wording or number, a lack of contextual sentences, an unclear table column caption, the irrelevant features in the bar and axis of the bar graph, and the unclear question wording. Those issues serve as empirical evidence on how the items could be refined regarding the wording in the text, the graphical and tabular displays, and the question wording. This refinement ensures that all items could assess the intended SL skills, and the item features are clear from the perspective of students. In addition, conducting a pilot study by using cognitive interviews enables researchers to determine whether the test items can facilitate various responses and avoid inaccurate responses brought on by instrument errors, resulting in valid test items for an SL test. However, the items resulting from this study were open-ended and intended for small-scale assessment. For large-scale assessments, it may not be effective to use cognitive interviews, as those in this study, to validate a large number of open-ended problems. Therefore, further study could focus on this limitation so that large-scale assessments can be conducted.

ACKNOWLEDGMENTS

The author would like to thank Tom Lowrie, Sitti Patahuddin, and Tracy Logan for supervising the author during his PhD. The author was an Australia Awards student and as such this study was conducted with the support of the Australia Awards Scholarship.

REFERENCES

- 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>
- 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>
- 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. https://doi.org/10.1007/1-4020-2278-6_1
- 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>
- Büscher, C. (2022). 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>
- 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>
- 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>
- 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>
- 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>
- 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
- 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>
- 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>
- 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>
- 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>
- Koga, S. (2022). 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>

- Kurnia, A. B., Lowrie, T., & Patahuddin, S. M. (2024). The development of high school students' statistical literacy across grade level. *Mathematics Education Research Journal*, 36(Suppl 1), 7-35. <https://doi.org/10.1007/s13394-023-00449-x>
- 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
- 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
- OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. OECD Publishing.
- 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. (2023). *PISA 2022 results (volume I): The state of learning and equity in education*. OECD Publishing. <https://doi.org/10.1787/53f23881-en>
- 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>
- Reinhart, A., Evans, C., Luby, A., Orellana, J., Meyer, M., Wiecezorek, 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>
- 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>
- Schild, M. (2017). GAISE 2016 promotes statistical literacy. *Statistics Education Research Journal*, 16(1), 50-54. <http://iase-web.org/Publications.php?p=SERJ>
- 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. (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. (2013). 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>
- 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>
- Tiruneh, D. T., De Cock, M., Weldelessie, 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>
- 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>
- 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>
- Weathington, B. L., Cunningham, C. J. & Pittenger, D. J. (2010). *Research methods for the behavioral and social sciences*. Wiley.
- 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.
- 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>