MOJES MALAYSIAN ONLINE JOURNAL OF EDUCATIONAL SCIENCES OKTOBER 2022, 10 (4)

PRE-SERVICE TEACHERS' PERCEPTIONS OF OUTLIERS IN STATISTICAL GRAPHS

Norabiatul Adawiah binti Abd Wahid
***Suzieleez Syrene binti Abdul Rahim**
Sharifah Norul Akmar binti Syed Zamri
Department of Mathematics and Science Education
Faculty of Education, Universiti Malaya
suzieleez@um.edu.my

ABSTRACT

Statistical graphs are often used by mass media to convey information to the community. Hence, it is a requirement for pupils in school to understand and apply various statistical skills to help them interpret the statistical graph. Teachers must play a very important role in ensuring that the student understands and are able to interpret various types of statistical graphs by applying an appropriate statistical skill. However, an outlier in a set of data always causes problems to the graph reader. An outlier can cause serious problems in statistical analyses. An outlier lies at an abnormal distance from other values in a random sample from a population. Therefore, this study looked at how prospective Mathematics teachers interpreted and build their perceptions towards an outlier in two types of statistical graphs by using an in-depth interview with all of the participants of this study. There were two types of statistical graphs used in this study, which are frequency table and a bar chart. Based on the findings obtained from this study, five prospective teachers involved in this study agreed that the outlier had various effects on the distribution of data for the entire data represented by the frequency table. This issue will be discussed in detail in this report. Hopefully, the findings from this study will give a clear description on how far the participants of this study perceived the role and effects of outliers in a set of data.

Keywords: *Outlier, Data, Statistical Graphs, Frequency Table, Bar Chart*

INTRODUCTION

Visual representations such as graphs and charts are important tools to make data and models more understandable. In our visual world, information is often presented in graphical format through diagrams, maps, and visual representation of data. Understanding mathematical relationships that is expressed in graphical form has been identified as a critical component of mathematics proficiency (Rosenblum et al., 2018). In our daily life, we are confronted with a lot of statistical graphs that visualize information, such as election results or the development of stock prices. It became more important in this pandemic era as the information about COVID-19 is mainly represented by various types of statistical graphs. Therefore, skills to understand, interpret and perceive the data set represented by those statistical graphs is really in need.

The study on graph comprehension was focused when Curcio (1987) introduced three levels of graph comprehension. Regardless of the graph form used (pie, line, bar, etc.), the three levels of graph comprehension are reading the data, reading between the data, and reading beyond the data**.** Even

though Curcio (1987) have given us ideas on how graph comprehension can be classified, we can't neglect the importance of statistical literacy in those processes. These aspects have been studied in detail by various researchers. Gal (2002)'s study which focuses on statistical literacy, has introduced the conceptual model of statistical literacy. In this model, Gal (2002) have divided statistical literacy into a building block as shown in Figure 1.

Knowledge elements	**Dispositional elements**
Literacy skills **Statistical knowledge** **Mathematical knowledge** **Context knowledge** **Critical questions**	**Beliefs and attitudes** **Critical stance**
Statistical literacy	

Figure 1. Conceptual Model of Statistical Literacy (Gal, 2002)

In Figure 1, it is clearly shown that there were three types of knowledge needed to support statistical literacy. These knowledges were statistical knowledge, mathematical knowledge and context knowledge. Among these three types of knowledge, context knowledge is something that have not been clearly thought when we learn about statistics and also mathematics at school. These aspects have brought many confusions and problems about the actual definition and needs of statistical literacy by various researchers. According to Gal (2019), he explains that instead of only depending on the knowledge that have been learned in the classroom, statistical information in data form is also needed to be related to the real-world situations. In relation with statistical graph comprehension, context knowledge is really needed. It cannot be denied that each data set that were presented by any types of statistical graphs has its own context. It depends on the graph reader on how they want to interpret those contexts, but the important point here is the presence and needs of contextual knowledge to allow the graph reader to interpret the statistical graphs appropriately.

Next, the idea about the levels in graph comprehension (Curcio, 1987) and statistical literacy by Gal (2002) have been used by Watson and Callingham (2003) when they introduced their idea about six levels of statistical literacy hierarchy. Those six levels of statistical literacy hierarchy were idiosyncratic understanding (level 1), informal understanding (level 2), inconsistent understanding (level 3), consistent non-critical understanding (level 4), critical understanding (level 5), and critical-mathematical understanding (level 6). These hierarchical levels have been mainly used to explain the level of understanding of students' arguments in reading and interpreting data via critical thinking.

Even though the study on how graph comprehension and the appropriate knowledge needed to support the comprehension have been introduced, there were a lot of challenges and mistakes that still happen while interpreting statistical graphs with an outlier. The evidence of this problem can be found in the study conducted by Estrada and Batenero (2008) who found that 45 % of the participants in their study did not take outliers into account when computing averages (Hannigan, Gill, & Leavy, 2013). Both researchers have also found that outliers were one of the misconceptions founded among their participants, prospective primary teachers. Meanwhile, Shoughnessy have stressed that one of the needs in making pedagogical content move in statistics lesson would be to superimpose the providing actual raw distribution along with the summary of critical values, including an outlier (Rossman, 2013). To make things worse, the findings from the study of Jacobbe and Carvalho (2011) have shown that the teachers involved in their study showed no understanding of the effect of outliers on the mean (Koleza & Kontogianni, 2016).

Based on the previous research findings, it proved that an outlier in a set of data has always given some problems to the graph readers/participants of the study. At the same time, an outlier always causes serious problems in statistical analyses. The position of an outlier that lies at an abnormal

distance from other values in a random sample from a population could give effect to the interpretations and analysis of the data set. Therefore, we decide that a study about how an outlier would affect the participants interpretation and perceptions towards the statistical graphs need to be conducted.

Be aware of the various problems that arise and have been determined by previous researchers as well as the importance of a prospective teacher to see the perception towards the outliers in a data set, therefore, the objective of this research is to determine on how the participants of this study build up their perceptions towards the outliers in statistical graphs. There were two different types of statistical graphs used in this study which were frequency table and bar chart. To ensure research objectives can be achieved, those two statistical graphs have at least one outlier in the data set.

METHODOLOGY

An in-depth interview has been used to collect all the information needed to fulfil the objective of this study. All of the participants have been interviewed individually. This is to ensure that the process of interpretation and perception towards each of the statistical graphs can easily be determined. Besides that, this step is taken to ensure that the feedback given by each participant were not being affected by the feedback givenk by other participants. At the same time, this process allowed us to pose questions based on each participants' interpretations only, without being affected by the others feedback. This helps us to increase the validity of our data.

There were two different types of statistical graphs used in this study. Those two statistical graphs were frequency table and bar chart as shown in Figure 2 and Figure 3 below.

Type of waste	Decomposition period
Banana peel	1–3 years
Orange peel	1–3 years
Box	0.5 year
Chewing gum	20–25 year
Newspaper	A few days
Polystyrene cup	More than 100 years

Figure 2. Statistical Graph 1

Figure 3:

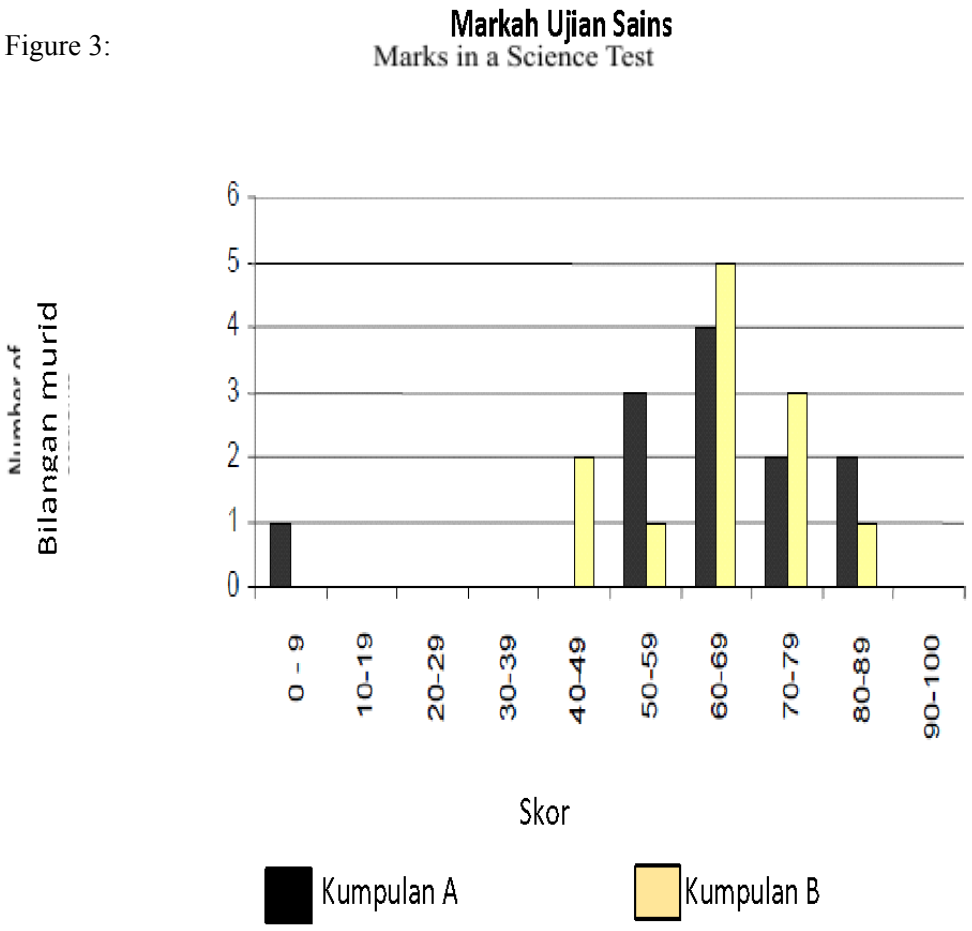

Figure 3. Statistical Graph 2

Those statistical graphs were given one after another to avoid confusion and to allow the participants to have a clear idea about the statistical graphs. By having the clear idea about the statistical graphs, hopefully it would help them to do their interpretation towards each statistical graph appropriately. The interview session started after the participants told us that he/she have clearly understood about the statistical graphs and when they were ready to discuss their interpretations and perceptions onto those statistical graphs. In accordance with the objective of this study, the findings that will be focused on the data discussion below will focus on the perceptions of all the participants to the outliers in those statistical graphs.

There were five participants who were involved in this study. The limited numbers of the participants was because this was a qualitative study. Besides that, we have conducted an in-depth interview with all of our participants separately between each of them. This is to ensure the reliability of all the data obtained from the interview session that have been conducted with all participants. Besides that, all of the participants would be asked based on their interpretations towards two types of statistical graphs, frequency table and bar chart. All of our participants were a pre-service teacher who were in their final year of their study at one of the Teachers Training Institute (IPG) in Malaysia.

DATA FINDINGS

The discussion on the data findings would be carried out separately. Therefore, the discussion on the perceptions towards the outliers in frequency table would be discussed first before the discussion on the perception towards the outlier of bar chart.

Perceptions Towards Outliers in Frequency Table

Frequency tables were always used as an organizer to a set of data. By using frequency table, it helps the reader of the data set to easily understand about all the information carried by the frequency table. At the same time, frequency table was usually used before analysis of the data set can be done in detail. There were similarities on how all of the participants start their discussion about the given frequency table. First, they would briefly discuss on the issue represented by the data set in the frequency table. All of the participants have agreed that the data set was about the decomposition period for five types of wastes, banana peels, orange peels, box, chewing gum, newspaper, and polystyrene cups. They have also discussed about the longest (polystyrene cup, more than 100 years) and less time (newspaper, a few days) of the decomposition period based on the data set. Based on this information, all of the participants have determined that the given data set have a very big range.

Based on that information, the participants have focused their discussion on the most suitable statistical graphs that can be used to represent the set of data. Based on the findings, its clearly shown that four participants faced some problems to determine the most suitable statistical graphs to represent the set of data. This is because the range between the data is too big that was brought by the two outliers in this data set. Because of this matter, there were some confusion when the participants discussed on what is the most suitable statistical graphs that can be used to represent the set of data.

To avoid problems while selecting the most suitable statistical graphs to represent the set of data, all of the participants agreed that those two outliers need to be taken out from the group. Even though the participants have suggested to take out those two outliers from the data set, but they still carry on their work in determining and drawing the most suitable statistical graphs for the data set by using the outliers. They have listed out line graphs and bar chart before they agree that pie chart is the best representation for the data set.

Four participants have suggested to use a pie chart to represent the set of data. But most of them decided to use pie chart after they found that their earlier suggestion is not suitable to represent the set of data. This happens when they got stuck while drawing their suggested statistical graphs. Because of this, they decided to change to use a statistical graph without any axis on it which is a pie chart as they realize that by having axis in the statistical graph and with a big range of data it would really affect the scale that needs to be used on their axis. Besides that, the big range would also result a large gap between each data in the data set.

The suggestion to use the pie chart to represent the set of data also gave various obstacles to the participants. They could not correctly build up the suggested pie chart. This happens when they calculate the angle of the segment for the outlier. Besides that, the big range in the data set has also given an effect to their calculations. By looking at their working steps, it can be clearly seen that there was one participant who clearly miscalculate the angle of segment, where she used 100% instead of 360° in her calculations. Towards the end of their calculations for the angle of each segment then they have decided there were no appropriate statistical graphs that can be used to represent this set of data. These can be clearly seen in Figure 4.

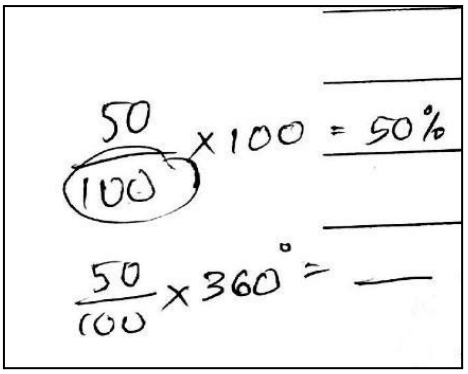

Figure 4. Miscalculation for the angle of segment in pie chart

Besides taking out the outliers from the data set, participants had also adjusted the value of the data in the data set. This is because some of the data were given in a range form that would affect the analysis of the whole data set. Therefore, all of the participants have adjusted the value of the data set as an example in Figure 5 that have been taken by one of the participants.

Jenis bahan buangan	Tempoh Penguraian
Kulit pisang	1–3 tahun 2 00
Kulit oren	1–3 tahun 2 00
Kotak	0.5 tahun 0.5 a
Gula-gula getah	20–25 tahun (25)
Akhbar	Beberapa hari 0.1 tahun
Cawan polisterine	Lebih 100 tahun 100

Figure 5. Adjustment to the value of data

When discussing about the purpose of doing so, the participant explained that by using this method it really helps her in drawing the appropriate statistical graphs for the data set.

Out of five participants in this study, there were one participant who has clearly shown his understanding towards the effect of the outlier in the given set of data while choosing the best and suitable statistical graphs to represent the data set. He has also suggested that line graph is the only solution if the data set still need to be represented by statistical graphs. In his discussion he also mentioned that the line graph that he suggested requires a lot of time to be drawn because of the effect brought by the outliers and the biggest range between the data.

Perceptions Towards Outliers in Bar Chart

The first thing that all of participants do after they received Figure 3 that contained the bar chart was to discuss on the issue represented by this statistical graph. Since the mean score for each group have been given, participants have focused their interpretations on the data set based on the central

tendency value and the data distribution between both of the groups. Based on information received from the interview session, it seems that all of the participants tend to focus their discussion on how the outlier of this set of data gave effect to the value of mean for a group of students. By looking at Figure 3, there were two different groups of students that have been compared on their achievements in a Science test. Even though the number of students in Group A that obtained 80-89 marks is more than Group B, there was a student from this group who only scored 0-9 marks in the same test. This difference has been discussed in detail by all of the participants in their interview session. Based on the findings that were obtained, all of the participants agreed that the only students from Group A who obtained 0-9 marks in this test can be classified as the outlier for the set of data for Group A.

All of the participants agreed that this outlier gave an effect to the mean for Group A in this test. Besides that, the outlier also gave an impact towards the value of median for this graph. By discussing the differences in the achievements between Group A and Group B, all of the participants agreed that the outliers in Group A have given an effect to the overall achievements of Group A compared to their friends in Group B. This aspect have been clearly discussed by the participants when they do the comparison by only looking at the central tendency value for each group. However, this might change in the analysis being done by making comparisons between the number of students in each group of marks. Besides comparing the achievements for each group, all of the participants have also discussed the reason behind the achievement of the outliers in Group A.

DISCUSSION

Based on the findings that have been obtained via an interview session with all of the participants, there were a few points of discussions that can be elaborated and compared with findings from the previous studies. Those points were as below:

a) Discussion about the context of the statistical graphs
b) Choice of the most suitable graphs based on the data set
c) The effect of outliers towards the value of central tendency

Discussion About Context of the Statistical Graphs

The importance of context in Mathematics study has been stressed by PISA when they include this aspect in their definition for numeracy. PISA defines numeracy as "an individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts," including "reasoning mathematically and using mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena" (Börner et al., 2019). In the study of statistics, Gal (2019) have stressed on the importance of context by listing out two contexts that were mainly found in the study of this field. Those two contexts were civic statistics and burning issues. Based on Figure 2, it brought us to an environmental context, which were categorized as burning issues, while Figure 2 brought us to educational achievement context which have been categorized as civic statistics.

By looking at how all of the participants began their discussion towards both statistical graphs, we can conclude that all of them have been able to determine and discuss their interpretations by using an appropriate context represented by both statistical graphs. Besides that, they have been able to make their conclusion about the data set correctly. The ability to correctly determine the context of both statistical graphs has helped them to elaborate their discussion on both statistical graphs. At the same time, this ability has helped them to discuss in detail about the data set including the effects of its outliers throughout their interview session. All participants have also shown that they were able to determine the exact role for each outlier to their data set. This shows that the participants have been able to use various statistical skills and knowledge throughout their interpretation and discussion on both statistical graphs. The importance of applying various skill sets has been stressed by Rosenblum et al. (2018) in their study. One of the participants in Rosenblum et al. (2018) study has stressed that an individual with good literacy skills is able to generalize the information from any statistical graphs. This information helped them to gain enough prior knowledge that was needed to support their overall literacy skills.

Choice of The Most Suitable Graphs Based On Data Set

In this study, we have determined that the participants have also discussed about the most suitable statistical graphs that can be used to represent the set by of data especially when they discussed about the frequency table. Based on our findings, it is clearly shown that most of the participants have faced problems to decide the most suitable statistical graphs to represent the set of data. This is mainly because of the presence of two outliers in the data set. These two outliers have resulted in a big range between the data in the data set. If we refer to Figure 2, it clearly showed that there were two different units of time involved, days and years. The biggest value in this frequency table is more than 100 years meanwhile the smallest value is only a few days. To avoid all of the problems brought by the big data range, most of the participants agreed that those two outliers need to be taken out from the group. Those two outliers were the smallest data, newspaper (few days), and the biggest data, polystyrene cup (more than 100 years).

The elimination of outliers from the data set shows that they have a good knowledge on how to manage the presence of outliers in a data set. According to Aguinis et al. (2013), there were many strategies that can be used while handling outliers in the data set including making some corrections to the data and also by removing the outlier. By referring to the work steps carried out by all of the participants when they handled those outliers in the data set, it clearly shows that those steps suggested by Aguinis et al. (2013) has also been carried out. Figure 5 showed us that the participants have tried to make some adjustments towards the data in the data set provided in Figure 2. This proved that the participants in this study have an appropriate knowledge on how to handle the outliers in the data set. However, four of the participants in this study still faced some problems in selecting the best statistical graphs to represent the given data set.

To determine the most appropriate statistical graphs for the data set, Börner et al. (2019) has determined that the participants in their study have conducted three steps. Those steps were (i) examine a graph and answer yes/no insight questions by modifying usage of graphic variable types; (ii) read a simple case study that defines an insight need and dataset, and then, select the best visualization, graphic symbols, and variable types to meet the predefined need; and (iii) listen to a client explaining a real-world problem, identify insight need(s), pick the most relevant dataset(s), construct an appropriate visualization, and verbally communicate key insights to the client. These steps have been categorized as scaffolding. By comparing these steps with the findings from our study, it shows that those processes have also been conducted by all of the participants. As mentioned above, the participants in this study have faced some challenges to determine the most suitable statistical graphs to represent the given data set. This problem is related to steps (ii) that have been suggested by Börner et al. (2019). Based on the findings in this study, it clearly shows that the problem in selecting the best statistical graphs to represent the data set have affected the (iii) steps that needed to be carried out by the participants. Even though Börner et al. (2019) have stressed the importance of the third steps to help on verbal communication of the key insights of any data set, but, it never happened in this study. The participants in this study were able to clearly communicate and discussed the key insights from the data set with the interviewer.

By comparing the data that was obtained with the levels of statistical literacy hierarchy (Watson & Callingham, 2003) we could categorize our participants was under Tier 2. Watson and Callingham (2003) have categorized Tier 1 as the ability to generally understand how to read data and explain data correctly. Ability in Tier 2 comprises analysis, interpretation, and making conclusions correctly accompanied by precise statistical terminology. Meanwhile in Tier 3, participants should be able to present data into other forms practically and be able to make predictions. Based on the findings, we have determined that four participants struggled to determine the most appropriate statistical graphs that can be used to represent the data set in Figure 2. According to Risqi and Setianingsih (2021) this showed that the participants need to have special treatment from their lecturer, especially in the context to improve statistical literacy skills.

The Effect of Outliers Towards the Value of Central Tendency

The presence of outliers in any data set could lead to biased estimation of central tendencies (Aguinis et al., 2013). The value of mean given in Figure 3 brought us to the possibility of variability in the variation of the data. In terms of our study, the ability of our participants to discuss in detail about the effect of outlier in similar data set proved that they have good knowledge and skills about statistics. This finding is contrasted with the findings by Boels and Bakker et al. (2019) in their review about conceptual difficulties when interpreting histograms. According to Boels and Bakker et al. (2019), the lack of understanding of the big ideas of data and distribution and how it effects the measures of centre and shape in different types of graphs have caused misinterpretation among their participants. Based on the findings from our study, we found that there were no misinterpretation happening among our participants when they discussed the bar graph given to them.

Besides that, the presence of outliers would bring out a fresh and meaningful perspective towards the data set. Based on our findings, we noticed that our participants have been able to use their statistical skills and knowledge to help them discuss and interpret about the effects brought by outliers to the value of central tendency and their data set.

CONCLUSION

Based on the findings of this study, we conclude that the outliers in a data set have given some effects on the participants perceptions towards the data set especially when they need to apply their critical skills by selecting the most appropriate statistical graphs for a set of data. By looking at the positive sides of our findings, the participants have been able to build up an appropriate perception towards the statistical skills and knowledge even though there were an outlier in the data set and there were also some challenges faced especially when they need to select the most suitable statistical graphs to represent the data set.

Hopefully this study provided an insightful idea on how the outliers affects the perceptions of our participants towards the data set. At the same time, the information gained from this study could be used by the lecturers of statistics at IPG to focus on enhancing the skills needed to help the pre-service teachers build up the correct perceptions towards data set with outliers.

However, more studies in the same field as this study need to be conducted to get a clearer picture on how teachers in Malaysia gave their perceptions towards any set of data with outliers in it. At the same time, this is highly important because the participants for this study is limited to five participants only.
Therefore, we suggest that for future research, there should be more studies focusing on the effects towards the outliers in a set of data. We should look at how outliers effect the comprehension on any types of statistical graphs. We cannot deny that outliers should give a big effect in data analysis, but how far it effects to various types of research participants either both students and teachers should be looked in detail in future studies.

REFERENCES

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, *16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2019). Conceptual difficulties when interpreting histograms: A review. *Educational Research Review*, *28*(September), 100291. https://doi.org/10.1016/j.edurev.2019.100291

Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(6), 1857–1864. https://doi.org/10.1073/pnas.1807180116

Curcio, F. R. (1987). Comprehension of Mathematical Relationships Expressed in Graphs. *Journal for Research in Mathematics Education*, *18*, 382–393.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. International Statistical Review, 70(1), 1-25.

Gal, I. (2019). Understanding statistical literacy: About knowledge of contexts and models. *Actas Del Tercer Congreso Internacional Virtual de Educación Estadística*, 1–15. Retrieved from http://digibug.ugr.es/bitstream/handle/10481/55029/gal.pdf?sequence=1&isAllowed=y

Hannigan, A., Gill, O., & Leavy, A. M. (2013). An investigation of prospective secondary mathematics teachers' conceptual knowledge of and attitudes towards statistics. *Journal of Mathematics Teacher Education*, *16*(6), 427–449. https://doi.org/10.1007/s10857-013-9246-3

Koleza, E., & Kontogianni, A. (2016). Statistics in primary education in Greece: How ready are primary teachers? In D. Ben-Zvi & K. Makar (Eds.), *The Teaching and Learning of Statistics: International Perspectives* (pp. 289–297). https://doi.org/10.1007/978-3-319-23470-0

Risqi, E. N., & Setianingsih, R. (2021). Statistical literacy of secondary school students in solving contextual problems taking into account the initial statistical ability. *Pi: Mathematics Education Journal*, *4*(1), 43–54.

Rosenblum, L. P., Cheng, L., & Beal, C. R. (2018). Teachers of students with visual impairments share experiences and advice for supporting students in understanding graphics. *Journal of Visual Impairment and Blindness*, *112*(5), 475–487. https://doi.org/10.1177/0145482X1811200505

Rossman, A. (2013). Interview with Mike Shaughnessy. *Journal of Statistics Education*, *21*(1), 1–27. Retrieved from http://www.amstat.org/publications/jse/v21n1/dunn/rossmanint.pdf

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, *2*(2), 3–46. Retrieved from http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(2)_Watson_Callingham.pdf