Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
SCI2021-2724

Students' Understanding of Outliers in the Context of Boxplot and Scatter Plot

Lukanda Kalobo
Senior Lecturer
Central University of Technology
South Africa

ATINER CONFERENCE PAPER SERIES No: SCI2021-2724

An Introduction to ATINER's Conference Paper Series

Conference papers are research/policy papers written and presented by academics at one of ATINER's academic events. ATINER's association started to publish this conference paper series in 2012. All published conference papers go through an initial peer review aiming at disseminating and improving the ideas expressed in each work. Authors welcome comments.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Kalobo, L. (2021). **"Students' Understanding of Outliers in the Context of Boxplot and Scatter Plot"**. Athens: ATINER's Conference Paper Series, No: **SCI2021-2724**.

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209 Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm
ISSN: 2241-2891
03/08/2021

Students' Understanding of Outliers in the Context of Boxplot and Scatter Plot

Lukanda Kalobo

Abstract

Identifying and dealing with outliers is one of the most important phases of data cleansing. By identifying and analysing outliers, users can either gain insight into abnormal patterns or purge the data of errors. It is common to consider boxplot and scatter plot in the identification of outliers. While these two procedures are useful for respectively determining univariate or bivariate data, it should be used cautiously, as understanding the characteristics of outliers is more important in identifying it. This article shows how, with appropriate guidance students can understand and identify outliers by using a boxplot or scatter plot. The purposive sample consisted of 61 university students who were enrolled in their first year course on Mathematics Education. For an hour, the students solved problems whose solutions required them to understand and identify univariate and bivariate outliers by using a boxplot and scatter plot. Upon quantitatively and qualitatively analysing the data, the students' statistical misunderstandings have been spotted, and classified in groups. Suggestions on how to confront these misunderstandings have been projected. We recommend that educators consider these statistical misinterpretations as they determine whether, how, and when to identify outliers using a boxplot or scatter plot.

Keywords: univariate, bivariate, outlier identification

Acknowledgments: The author would like to thank Ms MD, Kalobo, Mr MKS Kalobo and Mrs K. Kalobo for the support during the preparation of this paper.

Introduction

Representations play a crucial role as students build and improve their conceptual understanding in statistics. Data analysis as addressed in school mathematics curricula involves the use of such representations as boxplots, dot plots, scatterplots, stem and leaf plots, bar graphs, histograms, and so on as part of the tool kit for analyzing data (Edwards, Özgün-Koca & Barr, 2017). It is important to note a shift in current research from earlier research on understanding graphical representations. These earlier studies focused more on "graph sense", "graphicacy" and "graphical comprehension" than on distributional reasoning about graphs (Friel, Bright & Curcio, 1997; Friel, Curcio & Bright, 2001). Graphicacy is the ability to read and interpret graphs (Friel & Bright, 1996). Graph sense is the ability to recognize components of graphs, speak the language of graphs, understanding relationships between tables and graphs, respond to questions about graphs, recognize better graphs, and contextual awareness of graphs. Scientists often run into datasets that contain unusual values. An unusual value is a value which is well outside the usual norm. The unusual values which do not follow the norm are called an outlier**.** Outliers present a particular challenge for analysis, and thus it becomes essential to identify, understand and treat these values (Cousineau & Chartier, 2010). Identifying outliers is a significant problem that has been studied in various research and application areas (Wang, Bah, & Hammad, 2019). Nowadays, outlier identification is primarily studied as an independent knowledge discovery process merely because outliers might be indicators of interesting events that have never been known before (Ilango, Subramanian, & Vasudevan, 2012). A variety of outlier identification techniques have been developed in several research communities. Many of these techniques have been specifically developed for certain application domains, while others are more generic. Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is univariate or multivariate, usually just bivariate (Komorowski, Marshall, Salciccioli, & Crutain, 2016).

In South Africa, according to Department of Basic Education (2011), learners in the Further Education and Training (FET) band (Grade 10-12) are required to represent data effectively, choosing appropriately from: bar and compound bar graphs; histograms (grouped data); frequency polygons; pie charts; line and broken line graphs; ogives; box-and-whisker-plots; differentiate between symmetrical and skewed data and make relevant deductions; represent bivariate numerical data as a scatter plot and suggest intuitively whether a linear, quadratic or exponential function would best fit the data (Department of Basic Education, 2011; Laridon et al., 2004). Furthermore, learners in the Further Education and Training (FET) band (Grade 10-12) are required to identify outliers in the context of a scatter plot as well as the boxplot and whisker diagrams (Department of Basic Education, 2011).

The Purpose and Research Question

To investigate students' understanding of outliers in the context of boxplot and scatter plot, the following question was formulated:

> How do(can) students' understand and identify the outliers using a boxplot or scatter plot?

The answer to this investigation is part of the strategies to be used by mathematics teachers at High school to enhance understanding and the identification of outliers using a boxplot or scatter plot.

The Theoretical Framework of the Study

This study used a framework based on the work of Tukey (1977) on exploratory data analysis in which data are explored with graphical techniques. Exploratory data analysis focuses on the detection of unanticipated patterns and trends in data of all types, whether randomly sampled or not (Bakker, Biehler, & Konold, 2004).

Literature Review

The purpose of this literature review is to understand and describe the types of outliers, the boxplot, the scatter plot, and the methods for identifying the outliers. In order to understand and describe these concepts one must look at their definitions, how to represent the boxplot, how to interpret the scatters plot, and the identification of univariate and bivariate outliers.

Outliers Data

Outliers are observations or measures that are suspicious because they are much smaller or much larger than the vast majority of the observations (Cousineau & Chartier, 2010). Jindal, Panda, and Lavanya (2019) defined an outlier as a data point which is anomalous from the rest of the data based on some measure. In addition, an outlier in a data set is a value that is far away from the rest of the values in the data set (Jenkin, van Zyl, & Scheffler, 2012). Such a point often contains useful information about the loopholes of the system described by the data and hence is a major field of research for the data analyst. Jindal, Panda, and Lavanya (2019) distinguished three mutually exclusive types of outliers: Point outliers, contextual outliers, and collective outliers.

Point outliers – When a set of values is considered, outlier concerning most observations in a feature, we call it as point outlier; also, sometimes

termed as the univariate outlier. Jindal, Panda, and Lavanya (2019) explained that it is one of the easiest outliers to identify. This type of outliers is completely anomalous from the other data points are termed as Point Outlier.

Contextual outliers – A value being considered unusual given a specific context. For example, a temperature reading of 32 degrees in a day in July in London will be considered too unusual. However, the same temperature in Bengaluru will not be considered unusual (Jindal, Panda, & Lavanya, 2019).

Collective outliers – A group of observations appearing close to each other because of their similar values. According to Jindal, Panda, and Lavanya (2019) this type of outlier data is very difficult to find since the data may be distributed all over the dataset and is extremely difficult to group these and can be time consuming and cumbersome.

Methods for Identifying Outliers

There are four basic ways of handling outliers: Accommodation, Incorporation, Identification and Rejection (Barnett, 1978). This paper focused on the identification of outliers as it is recommended in the Mathematics Curriculum in South Africa (Department of Basic Education, 2011). Given that the outliers are data points lying far away from the majority of other data points, outliers in the data that is not normally distributed do not require identification. One difficulty with treatments of outliers is that there is no unanimously accepted theoretical framework for the treatment of outliers. Various fields have developed various methods and rare are the approaches that can be formulated with the concepts of another approach (Cousineau & Chartier, 2010). In this literature review, we will distinguish univariate outliers from multivariate outliers.

The Univariate Outliers Methods

The univariate methods include the Tukey Method, Histogram, and the boxplot. The South African, Mathematics Curriculum uses the Tukey method and the boxplots to identify the outliers (Department of Basic Education, 2011). Within the univariate cases, we will examine both the Tukey method and the boxplots where the population of scores is of an unknown distribution with a notable asymmetry (skewness).

Tukey Method

This method uses interquartile range to identify the outliers. The formula here is independent of mean, or standard deviation thus is not influenced by the extreme value (Tukey, 1977).

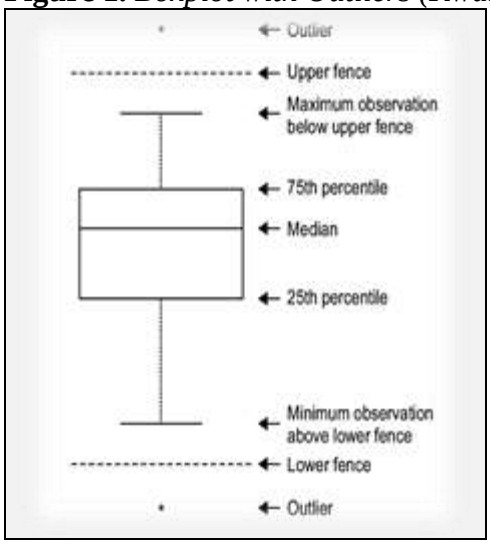
- Outlier on the upper side = 3rd Quartile + 1.5 * IQR; Outlier on the lower side = 1st Quartile – 1.5 * IQR; IQR (interquartile range) = 3rd Quartile – 1st Quartile (Kwak & Kim, 2017).

Boxplots

Boxplots are a powerful display for comparing distributions. They provide a compact view of where the data are cantered and how they are distributed over the range of the variable. In the USA, students learn about boxplots as young as about age 12. In some other countries, boxplots are introduced at a somewhat later age or not at all: in New Zealand to 13-14-year-olds, in Australia, Belgium, in Germany, the Netherlands, and South Africa to 15-16-year olds, and in France to 16-17-year-olds. In China and Israel, boxplots are not in the secondary school curriculum. Despite all the advantages in using boxplots to analyse data, we think there are several features of boxplots that pose particular challenges to students. Boxplots may not be difficult to construct. However, they are not so easy to understand, interpret, compare with one another, or match with other statistical representations of the same data (Edwards, Özgün-Koca & Barr, 2017; Lem, Onghena, Verschaffel, & Van Dooren, 2013a, 2013b).

The boxplots use the interquartile method with fences to find outliers, Boxplot is an excellent way of representing the statistical information about the median, third quartile, first quartile, and outlier bounds. Boxplots also called box-and-whisker plots or box-whisker plots give a good graphical image of the concentration of the data. The other name for boxplot is Tukey boxplots. A boxplot is constructed from the five-number summary (the minimum value, the first quartile, the median, the third quartile, and the maximum value) and, if there are outliers, the fences (Lemieux, 2017). The plot consists of a box representing values falling between IQR. The horizontal line inside the pot represents the median. The ends of vertical lines which extend from the box have horizontal lines at both ends are called as whiskers. Any value beyond these lines is called an outlier and is generally represented by discs (Figure 1).

Figure 1. *Boxplot with Outliers (Kwak & Kim, 2017)*



The identification of outliers in the observed distribution of a single variable spans the entire history of outlier identification not only because it is

the simplest formulation of the problem, but also because it is deceptively simple (Wilkinson, 2016). In Figure 1, the upper and lower fences represent values more and less than 75th and 25th percentiles (3rd and 1st quartiles), respectively, by 1.5 times the difference between the 3rd and 1st quartiles. Thus, any data item that is less than Q1 – 1.5 × IQR OR more than Q3 +1.5 × IQR is an outlier (Lehohla, 2013). An outlier is defined as the value above or below the upper or lower fences. In the boxplot in Figure 1, any data that lies outside the upper or lower fence lines is considered outliers. There are two different degrees of outliers. The mild outliers that lie beyond an inner fence and extreme outlier which are beyond an outer fence (Kwak & Kim, 2017).

The following example illustrates that in a box and whisker diagram, outliers are usually close to the whiskers of the diagram. This is because the centre of the diagram represents the data between the first and third quartiles, which is where 50% of the data lie, while the whiskers represent the extremes, the minimum and maximum of the data (Jenkin, van Zyl, & Scheffler, 2012). In the example below you are asked to find the outliers by drawing a box and whisker diagram and locating the data values on the diagram

0,5; 1; 1,1; 1,4; 2,4; 2,8; 3,5; 5,1; 5,2; 6; 6,5; 9,5

The first step is to determine the number summary. The minimum of the data set is 0.5. The maximum of the data set is 9.5. Since there are 12 values in the data set, the median lies between the sixth and seventh values, making it equal to $\frac{2.8+3.5}{2} = 3.15$. The first quartile lies between the third and fourth values, making it equal to $\frac{1.1+1.4}{2} = 1.25$. The third quartile lies between the ninth and tenth values, making it equal to $\frac{5.2+6}{2} = 5.6$. In the second step the box and whisker diagram are drawn (Figure 2). In this Figure 2, each value in the data set is shown with a black dot.

Figure 2. *The Box and Whisker Diagram (Jenkin, van Zyl, & Scheffler, 2012)*

The final step is about finding the outliers. From the diagram, in Figure 2, most the values are between 1 and 6. The only value that is very far away from this range is the maximum 9.5. Therefore, this is the only outlier in the data set. This is supported by Lehohla (2013), who defined an outlier as a data entry that is far removed from the other entries in the data set e.g., a data entry that is much smaller or much larger than the rest of the data values.

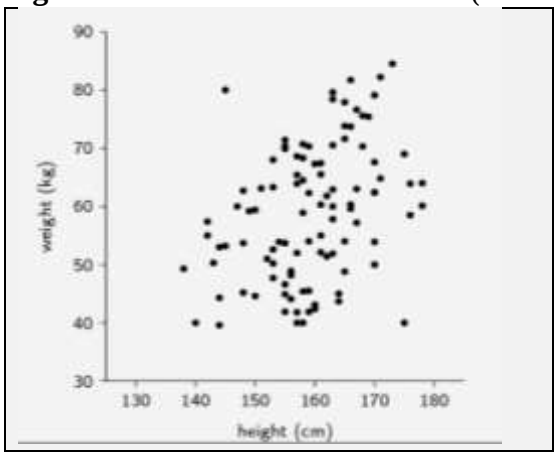
Multivariate Outliers Methods

Most of the outliers discussed in the above section are univariate outliers. We look at a data distribution for a single variable and find values that fall outside the distribution. Within the multivariate cases, we will consider the situation where the population is assumed to have two variables. However, we can use a scatterplot to identify outliers in a multivariate setting. In South Africa, the CAPS document stipulates that the Grade 11 learners do not need to learn how to draw the scatter plots, but they should be able to identify outliers on them (Department of Basic Education, 2011).

Scatter Plots

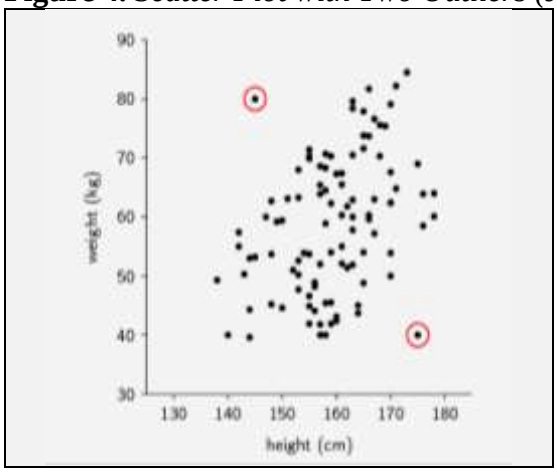
A scatter plot is a graph that shows the relationship between two random variables. We call these data bivariate (literally meaning two variables), and we plot the data for two different variables on one set of axes (Pagano, 2010). A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data (Triola, 2011). The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis (Pagano, 2010; Triola, 2011). The following example shows what a typical scatter plot looks like. In this example we have a data set that relates the heights and weights of a number of people. The height is the first variable, and its value is plotted along the horizontal axis. The weight is the second variable, and its value is plotted along the vertical axis. The data values are shown on the plot below (Figure 3).

Figure 3. *Scatter Plot with Outliers (Jenkin, van Zyl, & Scheffler, 2012)*



We can identify any outliers on the scatter plot. The plot in Figure 4 is visually inspected, it shows that there are two points that lie far away from the main data distribution. These two points are circled in the plot below (Figure 4).

Figure 4. *Scatter Plot with Two Outliers (Jenkin, van Zyl, & Scheffler, 2012)*



In the understanding and the identification of outliers in the context of boxplot and scatter plot, the student should be able to understand and use the boxplots and the scatter plots to identify outliers; demonstrate an awareness of the idea that the general pattern of a set of data, in terms of location, dispersion and skewness, can be graphically represented in a boxplot; understand that boxplots can be used to provide a quick and simple comparison of data sets.

Methodology

Sample and Sampling Technique

The target population (Lumadi, 2015; McMillan & Schumacher, 2014) of this study consists of all the first year Mathematics Education students at one of the University of Technology in South Africa. There were 61 students who participated out of 83 first year Mathematics students. A convenience sampling technique was used to select the first year Mathematics students the University of Technology according to their availability and the speediness with which data could be gathered (Lumadi, 2015).

Research Instrument

The data for this study were drawn from a student test consisting of four questions. The first three question involved problems on scatter plots and the last question concerned the scatter plot. Core issues related to the identification of the outliers using the boxplot and the scatter plot were used to build the students' test. For the boxplot, the core issues involved students' abilities to represent the statistical information about the maximum, minimum, median, third quartile, first quartile, outlier bounds and display the data using the boxplot. The core issues in scatter plot focused on students' abilities to make sense of statistical information, make interpretations, and identify the outliers.

Validity and Reliability

The content validity was strengthened by involving two academic mathematics educators in scrutinising the scope and depth of the questions (Fraenkel, Wallen, & Hyun, 2012). The data were analysed using descriptive statistics and the thematic analysis. In this study the findings cannot be generalised all the first year at all the University of Technology in South Africa because of the small size of the sample (Tipton, Hallberg, & Hedges, 2016).

Ethic

Confidentiality of the data and freedom to withdraw at any time without penalty were guaranteed to participants. An information sheet, explaining the purpose of the study, was given to all participants. Informed consent was obtained from students involved in this research project (McMillan & Schumacher, 2014).

Data Analysis Methods

Descriptive statistics in the form of frequency percentages were used to summarise the data collected from problem solving questions. Thematic analysis was used to analyse opened handed questions.

Results

Data analysis was accomplished through quantitative and qualitative analysis of 61 participants' responses to the test. The quantitative results speak to whether participants succeeded in identifying the outlies using the scatter plot and selecting the correct boxplot. Subsequently to this, a qualitative analysis is presented to identify participants' reasons for identifying the outlier(s) in the scatter plot and in the boxplot. In the following sections the students are referred to as participants 01, participants 02, …, and participant 61.

Scatter Plots

Participants' Performance in Identifying the Outliers

The researcher analysed participants performances in identifying the outliers on the scatter plot in Table 1 using question 1.1, question 2.1 and question 3.1. The results for each question in Table 1 are represented descriptively as frequencies and percentages. The responses to these questions (Table 1) indicate to the researcher how participants identify outliers.

Table 1. *Participants' Test Results (N=61)*

Result	Question 1.1	Question 2.1	Question 3.1	Question 4.1
Pass	98.4% (60)	95.1% (58)	37.7% (23)	0% (0)
Fail	1.6% (1)	4.9% (3)	62.3% (38)	100% (61)

From analysis of responses, it is clear from Table 1that 98.4% of participants were able to identify the outliers in question 1.1, while 1.6% of participants failed the question. The same happened in question 2.1, where 95.1% of participants were able to identify properly the outliers, whereas 4.9% of participants were unable. It is alarming that, the responses to question 3.1, indicate that 37.7% of participants succeeded in identifying the outliers, although 62.3% failed. This shows that in general participants have an idea on how to identify the outliers on the scatter plot, but their understanding of outliers might be a challenge. Thus, analysing the participants' reasons for identifying the outliers is crucial.

Participants' Reasons for Identifying the Outliers

Question 1.2, 2.2 and 3.2 were open ended questions and it aimed at capturing responses on participants' justifications for the identification of outliers in the scatter plots.

The participants' responses in question 1.2 were coded according to the following categories, namely: *Sharon is much larger/Brad much smaller; far from/Distant from points; Sharon weighs a lot more/Brad weighs a less; and two points differs*.

In terms of the category '*far from*/ *distant from'* participants expressed the following views. Participant 14 stated: 'Because those two points are far from the regression line'/Participant 59: 'These two points may be considered as outliers because they are far from the general trend of the other points. Sharon is too high from the rest of the other learners and Brad is too low'. Relating to category '*Sharon weighs a lot more/Brad weighs a less*' the following was said: Participant 19: 'Sharon's backpack weighs a lot more than those of other students in her weight group while Brad's backpack weighs much less than those of students in his weight group'/Participant 37: 'Because Sharon backpack weighs more, and Brad backpack weighs less compared to other students backpack weight'. Pertaining to category '*Sharon is much larger/Brad is much smaller'* participant answered as follow, Participant 4: 'Both points do not follow the general trend of the data point Sharon is much larger than the rest of the values data and point Brad is much smaller than the rest of the values in the data'. Regarding the category *'differ significantly'* participants hold the following opinions: Participant 45: 'Because, these points are differs significantly from other observation'. This shows that in overall students were able to give the correct justification for the identification of outliers in the scatter plot.

In the question 2.2, the participants' responses were coded according to the following categories, namely: '*higher quality rating and far higher'*. Regarding

the category '*higher quality rating'* participants hold the following opinions: *Participant 4*: 'This is because computer B has a higher quality rating and it appears to be affordable, according to its pricing of 500 dollars'/Participant 19: 'It is quality rating is higher (80) than the rest of the computers in its price range which is between $500 to $1000'/Participant 61: 'Because point B has (80) which has the highest quality rating'. Considering the category '*far higher that'* the following were said: Participant 37: 'Because point B has far higher quality rating compared to other computers'/Participant 45: 'B, because Malema wants to buy a computer whose qualify rating is far higher and the point that indicate a higher quality is B in the graph'. Thus, most of the participants gave the correct reasons for identifying the scatter plot outliers.

The participants responses in question 3.2 were coded according to the following categories, namely: '*points are farthest/ far away and Very low/scored lower*'. Pertaining the category '*points are farthest/far away*' the following were mentioned: Participant 25: 'A, both points are the farthest from the regression line'/Participant 29: 'B, because it is the farthest and higher but similar participation rate'/Participant 45: 'C, because the points are far away from the other plotted points on the graph'. Regarding the category '*Very low/scored lower*' participants hold the following opinions: Participant 7: 'B & C, they have a very low percentage and also, they are far away from other points'/Participant 19: 'A, B, & C, these states scored lower than other states with similar participation rates'/Participant 53: 'A, these states scored lower than other states with similar participation rates'. Only participant 14 spoke of: 'None of the above'. This is a matter of concert participants were not able to list all the outliers giving the correct reasons.

Boxplots

Participants' Performance in Choosing the Correct Boxplot Diagram

In question 4.1 participants were asked to choose the correct boxplot diagram. This could be done only if the participants were able to draw the boxplot using the Tukey method. This method uses interquartile range to identify the outliers. It is disturbing that in overall participants were not able to choose the correct boxplot diagram. Thus, in drawing the boxplots there are several features that present particular challenges to participants.

Participants' Reasons for the Identification of Outliers

Question 4.2 was an open ended one and it aimed at capturing responses on identifying any outlier and giving reasons. The Answer is supposed to be '*Yes, there are two outliers, 40 and 42, because 40 and 42* $\notin$ *[12.75; 38.75]'.* Pertaining participants who responded with a '*No'* the following were revealed: Participant 17: 'B, no, there are no outliers because 12,75 and 38,75 are included in our minimum (15) and in our maximum (42)'/Participant 47: 'B,

no, because the is no value that is more than the upper limit or lesser than the lover limit'/Participant 54: 'B, no, because I had to draw representing five number summaries'. Considering participants who responded with a '*Yes*' the following were mentioned: Participant 21: 'A, yes, 42, does not fit in a range of [12.75; 38.75]'/Participant 23: 'C, yes, fifteen lies lower and 40. 42 lies higher in the boxplot'/Participant 33: 'D, yes, because the two upper points 40 and 42 are two far from the other points of the data'/Participant 35: 'D, yes, there two outliers 40 and 42; because they are higher that the upper quartile range'. Participant 40 elaborated that: 'C, yes, there are 3 potential outliers. The member with age 15 is too or the youngest in the team. Two members of 40 and 42 are old in the team'/Participant 51: 'D, yes, there is an outlier in the boxplot, 42 is far from other numbers on the boxplot. The lower outlier = 22.5−9.75 = 12.75; the upper outlier 29 + 9.75 = 38.75'. This shows that most of participants were able to calculate the upper and lower fences. But they struggled to interpret that any data that lies outside the upper or lower fence lines is considered outliers.

Discussion

With reference to the research question on how do students' understand and identify the outliers using a boxplot and scatter plot? One would have expected all the students to answer the test question 1.1, 2.1 and 3.1 on identifying the scatter plots outliers correctly given that respondents have completed high school. Most of the students responded the question 1.1 and 2.1 correctly but struggled with the question 3.1 (Table 1). This study found that students had difficulties in identifying and interpreting the outliers on the scatter plots, such as using a diagram of scatter plot to identify the outliers and interpreting the outlier by visualising the diagram. This finding can be compared to those of a study by Meletiou-Mavrotheris and Lee (2010), which found that learners have difficulties in tackling tasks involving group comparison and critical information presented graphically, especially simple reading and interpretation tasks.

In response to the question 1.2, 2.2, and 3.2 on student's justification for the identification of the scatter plot outliers, the results show that more students were able to give correct justifications for identifying the outliers in question 1.2 and 2.2 but thrashed with question 3.2. For example, in question 3.2 the scatter plot has 3 outliers (A, B and C). The Justification for identifying these outliers is that the states scored lower than other states with similar participation rates. Most the students could not give the correct justification but identified one or two correct outliers (A/B/Cor AB/BC/AC). Thus, most of students are not aware about how to identify the outliers in the scatter plot. Inadequate information about the identification of outliers can lead to a misunderstanding or misinterpretation on the part of the students who may be confused as to what exactly is required of them. This might be the case in this test where the students give a correct answer for question 1 and question 2 but

struggled completely with question 3. This might be a result of inadequate teacher explanations regarding how to determine the relationship between X and Y in a scatter plot. Once outlier is detected, it must be represented in understandable form. The representation can be in the visual form of graphical display (Eberle & Holder, 2007).

In question 4.1, we were also interested to know whether the participating students were able to draw a boxplot and select the correct boxplot amongst the listed diagrams. Results show that no participant was able to select the correct diagram suggesting that most of students do not understand the identification of outliers. It would have been interesting to know if these students were taught the same contents at high school given that literature in South Africa suggests that the National Curriculum and Assessment Policy Statement (CAPS) is a single, comprehensive, and concise policy document, for learning and teaching in South African schools for grade R-12. It comprises CAPS for each approved school subject such the Mathematics CAPS Grade 10-12 (Department of Basic Education, 2011).

In response to question 4.2 on students identifying the outliers and giving a reason, the results display that most students were able or unable to identify the outliers and give a correct justification. For example, the student responded that there are no outliers but at the same time the justification shows that there are two outliers on the boxplot. Students misinterpreted the boxplot owing to insufficient understanding of outlier identification as a result of inadequate teacher explanations regarding how to determine the outlier in a boxplot (Dawson, 2011). In the present study, learners had difficulty representing data on boxplots (i.e., constructing boxplot and indicating outliers on boxplot). This finding supports those of Meletiou-Mavrotheris and Lee's (2010) study, in which a noticeable proportion of students had difficulties understanding graphical representations and interpretation of graphs.

The findings of the present study support the researcher's earlier comments that it is important to note a shift in current research from earlier research on understanding graphical representations. Thus, it becomes essential for the teachers to develop students in identifying, understanding, and treating the outliers in the context of boxplots and scatter plots. Ben-Zvi and Garfield (2005) suggest the use of alternative assessment methods to better understand and document students learning. Moreover, Lancaster and Tishkovskaya (2010) suggest shifting the focus of Statistics curricula from mathematical calculations to tasks of a practical nature. Hence, students need to be provided with problems involving scatter plots and boxplots outliers embedded in different real-life contexts so that they exercise what they have learned in a variety of ways.

Conclusion and Recommendations

In this paper, we stressed various methods for identifying outliers during an analysis of univariate and bivariate data. The first step in outliers identification is

to find the outliers using the Tukey method and to plot the data using methods such as a boxplot; furthermore, to identify outliers on the scatter plots. Understanding the statistics concepts involved in identifying the outliers is crucial. We also underlined the problem resulting from the decision on how to manage outliers based on the research results. Lastly, we proposed some recommendations allowing to respond to the study.

References

Bakker, A., Biehler, R. & Konold, C. (2004). *Should Young Students Learn About Boxplots?* Curricular Development in Statistics Education.

Barnett, V. (1978). The Study of Outliers: Purpose and Model. *Journal of the Royal Statistical Society, Series C* (*Applied Statistics*), 27(3), 242-250.

Ben-Zvi, D. & Garfield, J. (2005). *The challenge of developing statistical literacy, reasoning, and thinking*. Kluwer Academic Publishers

Cousineau D., Chartier, S. (2010). Outliers identification and treatment: a review. *International Journal of Psychological Research*, 3 (1), 59-68.

Dawson, R. (2011). How Significant Is a Boxplot Outlier? *Journal of Statistics Education*, 19(2).

Department of Basic Education (2011). Curriculum and Assessment Policy Statement: Further Education and Training Phase, Grades 10-12. Pretoria: DBE

Eberle, W., & Holder, L. (2007). Anomaly detection in data represented as graphs. *Intell. Data Anal.*, 11(6), 663–689.

Edwards, T. G., Aslı Özgün-Koca, A., & Barr, J. (2017). Interpretations of Boxplots: Helping Middle School Students to Think Outside the Box, *Journal of Statistics Education*, 25:1, 21-28.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). How to Design and Evaluate Research in Education (Edisi Kedelapan ed.). (S. Kiefer, Penyunt.) New York City: McGraw-Hill Companies.

Friel, S. N., & Bright, G. W. (1996). *Building a Theory of Graphicacy: How Do Students Read Graphs?* Retrieved from ERIC database, (ED395277).

Friel, S. N., Curcio, F. R., and Bright, G. W. (1997). Understanding students' understanding of graphs. *Mathematics teaching in the Middle School,* 3, 224-227.

Friel, S. N., Curcio, F. R., and Bright, G. W. (2001). "Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications," *Journal for Research in Mathematics Education*, 32(2), 124–158.

Ilango, V., Subramanian, R., & Vasudevan, V. (2012). A Five Step Procedure for Outlier Analysis in Data Mining. *European Journal of Scientific Research*, 75(3), 327-339.

Jenkin, A., van Zyl, M., & Scheffler, C. (2012). Everything Maths. Grade 11 Mathematics Teachers' Guide. Jenkin, van Zyl, & Scheffler Education.

Jindal, A., Panda, N., & Lavanya, K. (2019). Comparison of Outlier Identification Techniques using KNIME Analytics Platform. *International Journal of Engineering Research & Technology,* 8(3).

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). *Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records* [Internet]. Cham (CH): Springer.

Kwak, S.K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, 70(4): 407.

Lancaster, G. A., & Tishkovskaya, S. (2010). *Teaching strategies to promote statistical literacy: Review and implementation*. ICOTS8 Contributed Paper.

Laridon, P., Barnes, H., Jawurek, A., Kitto, A., Myburgh, M., Pike, M., Myburgh, M., Rhodes-Houghton, R., Scheiber, J., Sigabi, M.& Wilson, H. (2004). *Classroom Mathematics. Grade 10 learners book*. Heinemann Publishers (Pty) Ltd.

Lehohla, P. (2013). *Data Handling and Probability* (Grades 10, 11 and 12). Statistics South

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013a). The Heuristic Interpretation of Boxplots. *Learning and Instruction*, 26, 22–35.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013b). On the Misinterpretation of Histograms and Boxplots. *Educational Psychology*, 33(2), 155–174.

Lemieux, C. (2017). *Measures of Location and Boxplots. How to Deal with Outliers in Your Data.* All things Data-Driven Marketing.

Lumadi M. W. (2015). The Logic of Sampling. In Okeke C. & Van Wyk M. (Eds.). *Educational Research: An African Approach.* Oxford University Press.

McMillan J., & Schumacher, S. (2014). *Research in education: Evidence-Based Inquiry*. Pearson.

Meletiou-Mavrotheris, M. & Lee, C. (2010). Investigating College-Level Introductory Statistics Students Prior Knowledge of Graphing. *Canadian Journal of Science, Mathematics and Technology Education*, 10(4), 339-355.

Pagano, R. R. (2010). *Understanding Statistics in the behavioural sciences*. 9th edition. Belmont, California: Wadsworth.

Tipton, E., Hallberg, K., & Hedges, L.V. (2016). Implications of Small Samples for Generalization: Adjustments and Rules of Thumb.

Triola, M. F. (2011). Essentials of Statistics. 4th edition. Boston, Massachusetts: Pearson Education.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in Outlier Detection Techniques: A Survey. IEEE Access 7: 1-1.

Wilkinson, L. (2016). *Visualizing Big Data Outliers through Distributed Aggregation*.