Developing box plots while navigating the maze of data representations

Bruce Duncan
University of Tasmania
<bruce.duncan@utas.edu.au>
Noleine Fitzallen
University of Tasmania
<noleine.fitzallen@utas.edu.au>

The learning sequence described in this article was developed to provide students with a demonstration of the development of box plots from authentic data as an illustration of the advantages gained from using multiple forms of data representation. The sequence follows an authentic process that starts with a problem to which data representations provide the solution. The advantage of using box plots is that they allow clear and efficient comparison of related data sets. In this case, students are given a maze on paper and timed while they complete it. This produces the first set of data. They then attempt the maze again, expecting that their time to do this will decrease. The need to compare these two data sets arises from the question, "Did the group improve their maze times on their second attempt?"

Background

The use of graphs in the mathematics classroom is first introduced in the *Australian Curriculum: Mathematics* in Year 2, when students are expected to create and interpret picture graphs. Column graphs are introduced in Year 3, dot plots in Year 5, stem-and-leaf displays in Year 7, histograms in Year 9, and box plots and scatter plots in Year 10 (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2013). The intention is that the introduction of different graphical representations be developmental and cumulative.

The thinking and reasoning required to interpret the different graphical representations increases as students progress through the compulsory years of schooling. Once a new graph type is introduced the expectation is that it will be used in future years, even though it may not be named explicitly in the content descriptions of the curriculum for the proceeding years. There is, however, very little attention given to the need to assist students to make connections among the various graphical representations in the curriculum.

The connections among the "range of graph types" and "multiple representations" are not acknowledged in the curriculum until Year 10 when students are expected to "Compare shapes of box plots to corresponding histograms and dot plots" (ACARA, 2013, p. 71). Year 10 graphing activities often include generating box plots from histograms. To be able to do this confidently it would be beneficial to provide students with the opportunity to establish an understanding of the relationship between different graphical representations earlier on in the curriculum and in contexts in which the purpose for representing data in different ways is made clear. It is appropriate to do so because younger students have demonstrated the ability to create and interpret scatter plots and box plots long before they are formally introduced in the curriculum (e.g., Cobb, McClain & Gravemeijer, 2003; Fitzallen, 2012; Özgün & Edwards, 2013).

The benefits of using box plots and scatter plots in classrooms prior to Year 10 are that students have the time to develop exploratory data analysis strategies and fundamental intuitions about working with data before focusing on the formal statistical interpretation of data using correlation coefficients for scatter plots and quartiles for box plots. Likewise, providing students with the opportunity to develop an understanding of the relationship between different graphical representations before Year 10 is beneficial.

Box plots

A box plot summarises a data set, locates the median, displays the spread and skewness of the data, as well as identifies the outliers, but does not display the overall distribution of the data (Friel, Curcio & Bright; 2001). A box is comprised of the interquartile range (IQR), which represents the middle 50% of the data. The IQR extends from the first quartile to the third quartile. Figure 1 shows that the IQR is divided by the median. The range of the left hand side of the IQR is smaller than the range of the right hand side of the IQR. This poses problems for some students because they find it difficult to understand that although there is the same number of data points represented in each section of a box plot, the size of each section is dependent on the density or spread of the data (Bakker, Biehler & Konold, 2005). This means that the data in the left hand side of the IQR in Figure 1 are closer together than in the right hand side of the IQR. Attached to the box on the left hand side and extending from the first quartile to the minimum value of the data set is a whisker, which represents the lower 25% of the data set. Another whisker is attached to the right hand side of the box. That whisker represents the upper 25% of the data set and extends from the third quartile to the maximum value. The whiskers obey the same principles of density and distribution as the box.

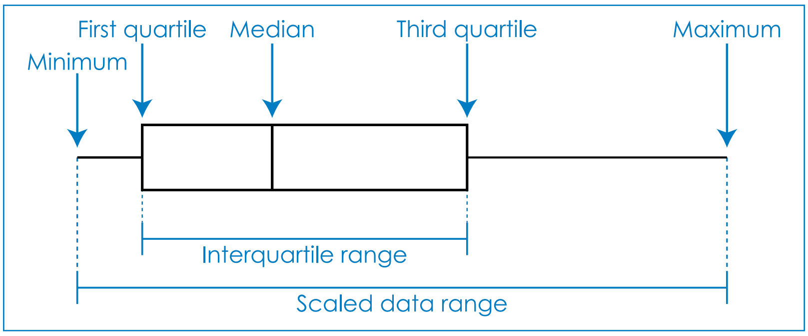


Figure 1. The structure of a box plot.

Although individual box plots are useful, box plots were developed for comparing multiple data sets (Tukey, 1977). Direct comparison of several data sets or subsets of data can be conducted efficiently by analysing the box plots displayed in parallel, as can be seen in Figure 2, which displays the data for the body weight of students from Years 1, 3, 5, and 7.

TinkerPlots: Dynamic Data Exploration (Konold & Miller, 2011) is a statistical software program that students can use easily to generate box plots (Watson,

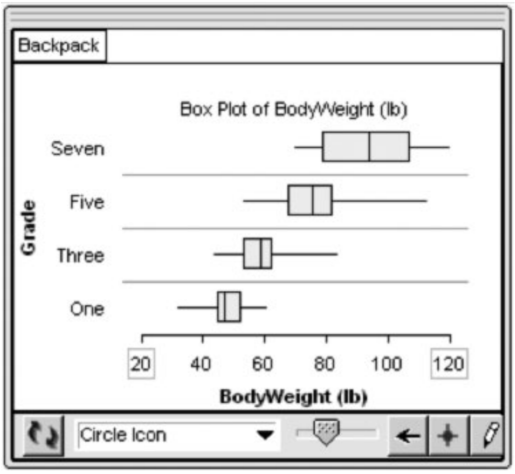


Figure 2. Parallel box-and-whisker plots produced using a data set in *TinkerPlots*.

Fitzallen, Wilson & Creed, 2008). Another useful technology tool for generating box plots is the CAS calculator, such as the TI-Nspire (Özgün & Edwards, 2013). The advantage of using the displays generated by these two options or similar technology innovations is that the data points can be displayed in conjunction with the box plot representation (Figure 3). Such displays allow students to see the direct connections among the distribution of the data and the corresponding parts of the box plot, thereby making links between the two graphical representations (Watson et al.). This enhances the opportunity for students to develop understanding of the purpose of each type of graphical representation as they interpret and make sense of the displays.

Stem-and-leaf displays

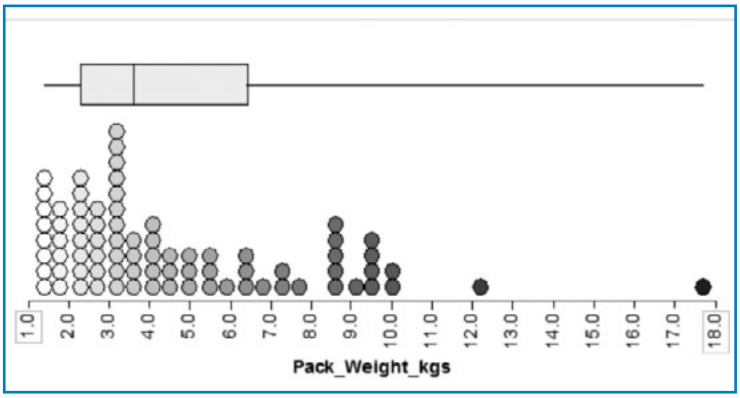


Figure 3. Box plots with data.

The stem-and-leaf display is an alternative to tallying values into frequency distributions. It organises a batch of numbers graphically and directs attention to various features of the data. It displays a distribution of a variable with the digits themselves making up the leaves of the display. The interval widths are displayed on a contracted number line, which makes up the stem of the display. Usually displayed vertically, it resembles a horizontal stacked dot plot (Figure 4). The development of stem-and-leaf displays should be understood as a way of representing the characteristics of the data set, while maintaining the identity of each datum. Groups are conserved and frequencies are clearly represented in the stem-and-leaf display, which can be seen as a sophisticated variation of the stacked dot plot.

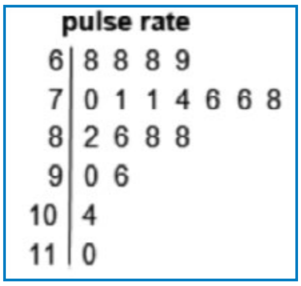


Figure 4. Stem-and-leaf display (ACARA, 2013, p. 122).

Stacked dot plots, like in Figure 3, provide a representation of frequency distribution that can be easily described. Because each datum is represented in relation to each other, although not explicitly, the characteristics of the data set are revealed. The distribution of two data sets can also be compared when displayed as a back-to-back stem-and-leaf display. This is demonstrated in Figure 5, which displays students' pulse rates before and after undertaking some exercise.

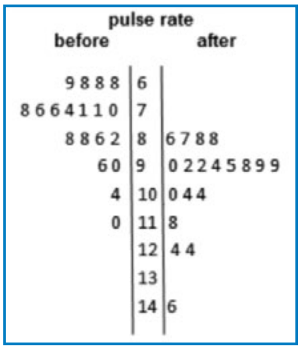


Figure 5. Example of back-to-back stem-and-leaf display (ACARA, 2013, p. 89).

Generating box plots from stem-and-leaf displays: The maze investigation

The Maze Investigation is an activity that provides students with the opportunity to answer the question: "Do people complete mazes faster the second time around?" To be able to answer the question there is a need to have two data sets to see if maze completion times improve if completed twice. The activity is run twice with students recording the time it takes for Trail 1 and Trial 2. The maze used to collect the data presented in this article can be downloaded from www.printablemazes.net (Figure 6).

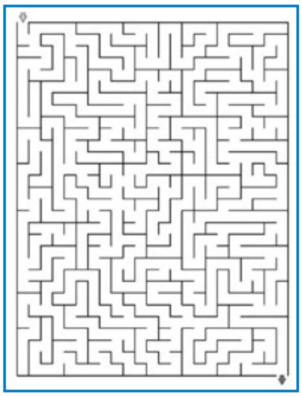
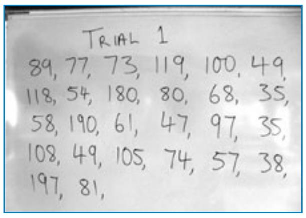
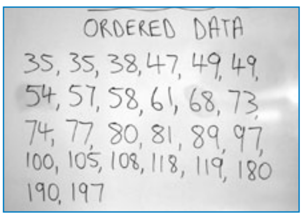
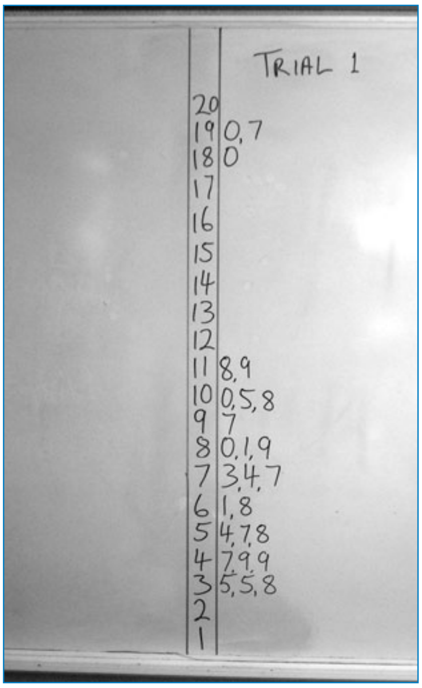
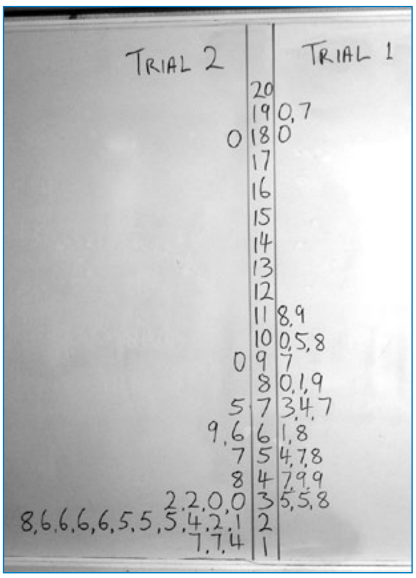
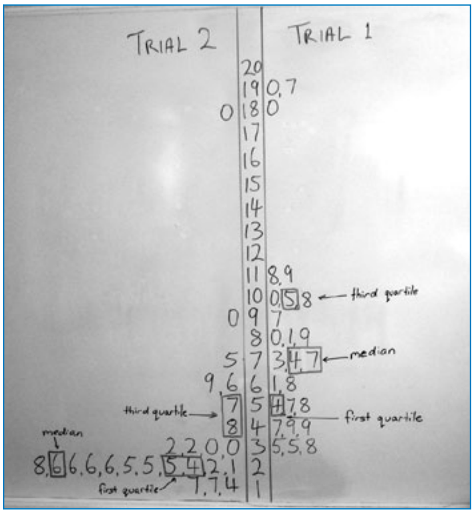
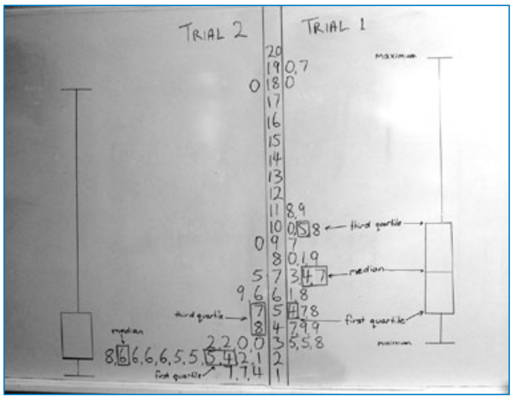


Figure 6. Maze #3, Set 5 (www.printablemazes.net).

Following social-constructivist pedagogy (Simon, 1995), the potential to develop students' understanding is increased when they themselves are required to determine the method by which the problem should be solved. For this activity, carefully scaffolded discussion can guide students from the raw data, through the process of analysis and representation, to the final representation that allows effective comparison between data sets. As each representation is developed, the discussion identifies the advantages gained by each progressive representation as a response to the question "Do people complete mazes faster the second time around?" is formulated. At the same time, the corresponding disadvantages that come from simplifying the representation should also be made explicit. The following activity sequence outlines the activity process and teaching opportunities that arise. The data for the worked example were generated by a class of adult learners.

<table>
<tr><th>Activity sequence</th><th>Description</th><th>Teaching opportunities</th></tr>
<tr><td>1. Posing the problem and identifying the question to be answered.</td><td>"Do people complete mazes faster the second time around?"</td><td>The process of data collection and representation is shown to have an authentic purpose.
Medium Mazes Set 5: Run-of-the-Mill (www.printablemazes.net)</td></tr>
<tr><td>2. The event.</td><td>Every student receives a copy of the maze, face down, and is instructed to turn the paper over and attempt the maze when the teacher says, "Go." The teacher starts a stopwatch on a data projector that all students can see and they attempt to complete the maze by drawing a path from start to finish without crossing any lines. When students finish they record the time on the stopwatch as the duration of their attempts.</td><td>The teacher may need to establish an upper limit for the duration of this task, by which time some students may not have finished.
Stopwatch (www.online-stopwatch.com/large-stopwatch/)</td></tr>
<tr><td>3. Raw data.</td><td>The time taken for each student to complete the maze is collected on a board at the front of the classroom. Initially, these data are collected in a random order to produce a list.

</td><td>The need to organise data can be made clear by first collecting data from students in a random order, such as "around the room."</td></tr>
<tr><td>4. Ordering data.</td><td>Students asked to consider, "How can we make these data easier to read?" and "How can we describe this set of results?"

</td><td>The advantage in ordering data can be made clear to students by scaffolding discussion about organising the data.</td></tr>
</table>

Activity sequence	Description	Teaching opportunities
5. Grouping.	Students discuss how the data can be grouped and then group the data according to a strategy selected by the class, which becomes the stem of the stem-and-leaf display.	The collected data can be seen as a sample from a possibly continuous range of measurements and that it therefore makes sense to speak of the frequency of outcomes within specified intervals (grouped data) rather than the frequency of occurrence of particular measurements.
6. Stem-and leaf-displays.	An appropriate scale is determined by discussion and drawn on the board and the data are recorded. 	Now the purpose of organising the data can be made clear through discussions that attempt to describe the data set by asking questions such as "What can we say about the data?" The data are analysed, organised, and represented in different ways to identify the range, any skewed distribution, and central tendency. The focus now shifts from students identifying their individual information to looking more broadly at the data from the whole group.
7. The second event.	The maze activity (step 2) is repeated with the same maze and times recorded.	Discussion should elicit the expectation that durations to complete the maze the second time around may become shorter. This comparison can be discussed informally after the data have been collected but before the data are organised so that the data are seen to confirm an explanation.
9. Organising the second set of data.	Students organise data from Trial 2 into a back-to-back stem-and-leaf display with the data from Trial 1. 	This process is a repetition of the process undertaken on the first data set. The opportunity exists, therefore, to allow students to carry out this process with greater independence from the teacher. In the example the data shows a very dramatic improvement in times, one that would be obvious from the raw data. A more challenging maze or a younger group of students may produce data that are less markedly different.

Activity sequence	Description	Teaching opportunities
10. Comparing data sets: Representations with a shared scale in a back-to-back stem-and-leaf display.	Students discuss the questions “How does this representation help us answer our question? Are the second times faster? Why do you say that?”	The closest comparisons can be made by comparing data sets on a common scale. Once again the discussion should be guided by the purpose so a good, guiding question here is, “How can we compare your maze completion time from Trial 1 with the completion time in Trial 2?” Discussion includes the comparison of the characteristics of each data set - range, skew, central tendency.
11. Medians and quartiles.	Students discuss “What is the middle score?” or “What score divides this group in half?”	Establishment of these features pre-empts the box plots but the discussion must focus students’ understanding on these terms as characteristics of the population, not the range. Once students understand that the median is determined by considering the number of scores in order, rather than the value of each score, the concept of quartiles, dividing the population into four equal sized groups, follows as a natural progression.
12. Box plots.	Students identify the five points on the stem-and-leaf display (minimum, first quartile, median, third quartile, maximum) and mark against the same scale to create the box plot. 	Box plots can be seen as simplified stem-and-leaf displays. Although the detail of each datum is lost, the simplification of this representation allows the data set to occupy less space and, therefore, makes box plots appropriate for the purpose of comparison. 
13. Answering the question.	“Do people complete mazes faster the second time around?” Attention can then be given to thinking about the informal inferences that can be made from the data, asking “Do you think another group of students would get the same result? Can we claim that students always complete Trial 2 quicker than Trial 1?	Comparison of the two box plots shows that the interquartile ranges do not overlap, therefore, the claim can be made that the people in the group were faster the second time round. Note that the first quartile and the median in Trial 2 fall at the same point on the vertical scale. That results in an unconventional looking box (interquartile range). Anomalies such as this arise when using real life data and present the opportunity to discuss why the representation looks different to what was expected.

Conclusion

By using a problem as a context for developing data representations the process is seen to be authentic. Maintaining students' involvement in that process by asking questions such as "How can we make this clearer?" illustrates not only the construction of the graphical representations but also the application of the properties of those representations. However, data collected from real life situations do not always result in a perfect example of the graphical representation developed. Although more challenging for teachers, it is worthwhile students exploring those data sets to develop the skills needed to be able to think flexibly when interpreting graphs. Although using contrived data sets that behave in a particular way may result in graphical representations that are simpler to explain, collecting data generated from an activity contributes to the authenticity of the learning experience.

References

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2013). *The Australian Curriculum: Mathematics, Version 5.1*, Monday, August 5, 2013. Retrieved from http://www.australiancurriculum.edu.au/Mathematics/Curriculum/F-10

Bakker, A., Biehler, R. & Konold, C. (2005). Should young students learn about box plots? In G. Burrill & M. Camden (Eds), *Curricular development in statistics education: International Association for Education (IASE) Roundtable* (pp. 163–173). The Netherlands: International Statistics Institute.

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction, 21*(1), 1–78.

Fitzallen, N. (2012). Students reasoning about covariation. In J. Dindyal, L. P. Cheng & S. F. Ng (Eds), *Mathematics education: Expanding horizons* (Proceedings of the 35th annual conference of the Mathematics Education Research Group of Australasia, Singapore). Sydney: MERGA.

Friel, S. N., Curcio, F. R. & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*, 124–158.

Konold, C. & Miller, C. (2012). *TinkerPlots: Dynamic data exploration, Version 2* [software]. Emeryville, CA: Key Curriculum Press.

Özgün, S. A. & Edwards, T. G. (2013). Interpreting box plots with multiple linked representations. *Mathematics Teaching in the Middle School, 18*(8), 508–511.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114–145.

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

Watson, J. M., Fitzallen, N. E., Wilson, K. G. & Creed, J. F. (2008). The representational value of hats. *Mathematics Teaching in the Middle School, 14*(1), 4–10.