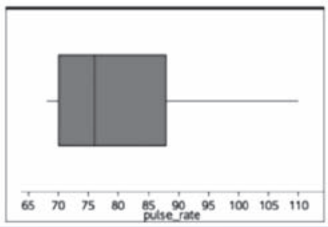
Box Plots
in the Australian Curriculum

Jane M. Watson
University of Tasmania
<jane.watson@utas.edu.au>

The purpose of this article is to compare the definition of *box plot* as used in the *Australian Curriculum: Mathematics* with other definitions used in the education community; to describe the difficulties students experience when dealing with box plots; and to discuss the elaboration that is necessary to enable teachers to develop the knowledge necessary to use them effectively.

The box plot is 40 years old (Tukey, 1970, 1977). Like many 40-year-olds the box plot has changed its shape and varied its personal descriptors over the years (Wickham & Stryjewski, 2011). Its power to inform exploratory data analysis has gained it a place in statistics curricula at various levels in many countries. The *Australian Curriculum: Mathematics* (Australian Curriculum, Assessment and Reporting Authority [ACARA], 2012) has placed it at Year 10 in the Statistics and Probability strand of the learning area. As the curriculum document is sparse in its description and elaboration of the term, three issues relating to the needs of the teachers required to implement the curriculum are discussed here: there is not one universally accepted definition of a box plot and teachers need to be aware that they and their students may encounter contradictions in different contexts where different conventions apply; teachers need to be aware of the difficulties their students encounter in understanding box plots; and, there is minimal advice in the curriculum elaborations about how the box plot can be used effectively in a statistical investigation.

The box-and-whisker plot below has been constructed from the five-number summary of the resting pulse rates of 17 students.



The term ‘box-and-whisker plot’ is commonly abbreviated to ‘box plot’.
A five-number-summary is a method for summarising a data set using five statistics, the minimum value, the lower quartile, the median, the upper quartile and the maximum value.

Figure 1. ACARA (2012) definition of Box plot.

Definitions

The extract from *Australian Curriculum: Mathematics* Glossary that defines box plot is shown in Figure 1. To assist in understanding the definition, the extracts for quartile and median are presented in Figure 2. Choosing a box plot format that extends the whiskers to the minimum and maximum values in the data set makes instructions for students straightforward but this is not the common convention in tertiary statistics courses. There the convention is usually to extend whiskers from the quartiles that form the edges of the box to the farthest values from the centre that are less than or equal to 1.5 times the interquartile range. In fact under the definition of parallel box plots, there is implicit acknowledgement of this convention as seen in Figure 3. MacGillivray (2011) in writing for the TIMES project supporting the *Australian Curriculum: Mathematics*, Year 10, Data Investigation and Interpretation, first acknowledges the definition of box plot in the curriculum and then immediately moves to the more common tertiary convention. This description is shown in Figure 4. If all data values lie within 1.5 times the interquartile range from their closest quartiles, then the box plots using the two conventions will look the same.

For the user of box plots the major difficulty with the definition of box plot in the curriculum is the ambiguity associated with the definition of quartile when the number of data values is not evenly divisible by 4. The use of “approximately” can lead to many possibilities for decisions on the placement of the edges of the box. A common convention is to use the median of the upper and lower halves of the data (historically called the Tukey hinge). Given that the median is defined (see Figure 2) this would appear to be a reasonable convention for teachers to use when an “approximate” value is required. The use of the word “approximately” in the definition of quartile means that in many instances there is unlikely to be one “correct” answer to the shape of a box plot, which in turn means that teachers will need to be flexible in marking student work.

The discussion so far has illustrated that unlike many other areas of the mathematics curriculum where definitions are fixed and immutable, the definition of box plot is not. Because students will meet box plots in many contexts in other learning areas and in the world outside of school, it is important to impress upon them the need for flexibility in interpreting new box plots presented to them. As Wall and Benson (2007) show, students are going to be reading and interpreting many different representations across learning areas as they progress through school. Wickham and Stryjewski

Quartile

Quartiles are the values that divide an ordered data set into four (approximately) equal parts. It is only possible to divide a data set into exactly four equal parts when the number of data of values is a multiple of four.

There are three quartiles. The first, the **lower quartile** (Q1) divides off (approximately) the lower 25% of data values. The second quartile (Q2) is the median. The third quartile, the **upper quartile** (Q3), divides off (approximately) the upper 25% of data values.

Percentiles are the values that divide an ordered data set into 100 (approximately) equal parts. It is only possible to divide a data set into exactly 100 equal parts when the number of data values is a multiple of one hundred.

There are 99 percentiles. Within the above limitations, the first percentile divides off the lower 1% of data values. The second, the lower 2% and so on. In particular, the **lower quartile** (Q1) is the 25th percentile, the **median** is the 50th percentile and the **upper quartile** is the 75th percentile.

Median

The **median** is the value in a set of ordered data that divides the data into two parts. It is frequently called the ‘middle value’.

Where the number of observations is **odd**, the median is the middle value.

For example, for the following ordered data set with an odd number of observations, the median value is five.

1 3 3 4 5 6 8 9 9

Where the number of observations is **even**, the median is calculated as the mean of the two central values.

For example, in the following ordered data set, the two central values are 5 and 6, and median value is the mean of these two values, 5.5

1 3 3 4 5 6 8 9 9 10

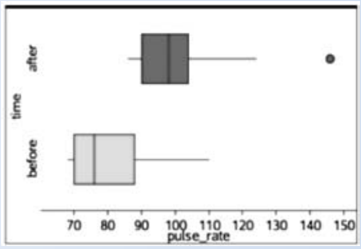
The median provides a measure of location of a data set that is suitable for both symmetric and skewed distributions and is also relatively insensitive to outliers.

Figure 2. ACARA (2012) definitions of Quartile and Median.

(2011) make this point even more strongly for box plots as they survey the development of the box plot over the last 40 years. A few of the adaptations they discuss include making the width of the plot proportional to the size of the data set for comparing data sets of different sizes; displaying density of values within the box using the shape of the box (narrower for fewer points and wider for more); using colour to indicate density with darker shading being more dense; and developing conventions for two-dimensional data sets. Although students might be shown some of these alternatives to create an awareness of diversity, it is unlikely they would be employed in the classroom.

Teachers themselves need to be aware of the variation of the presentation of system data in box plot form as part of their professional lives. The example shown in Figure 5 is of the type provided by the Victorian Curriculum and Assessment Authority (VCAA) (2010) for a hypothetical school's NAPLAN results. As can be seen from the key, 20% of the data are not represented in the box plots. For those not attending to the key, there may be a belief that the range of scores for the school for Writing is larger than the range for the state. As the school is a subset of the state, this is clearly impossible, regardless of the appearance of the box plot. Similar confusion could occur for the boxes in Figure 6 where there are no whiskers and 40% of the data are missing in plots reporting the outcomes for the five NAPLAN learning areas for two imaginary students. Pierce and Chick (in press) found that many teachers struggled with appropriate interpretation of plots like those in Figures 5 and 6.

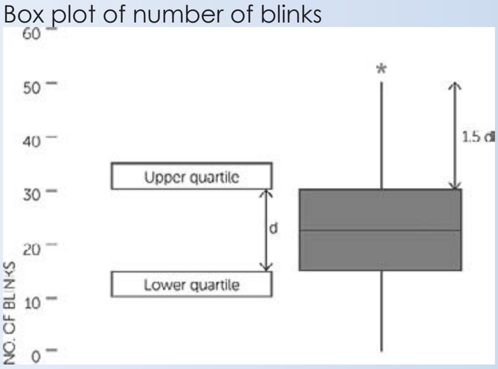
The message for students, and in many cases for teachers as well, is "Read the key" for any box plot encountered. The examples given here are evidence that there is no universally accepted convention for presenting summaries of data sets in box plot form. The power of box plots is in their ability to act as an intermediary between a complete plot of a data set, perhaps as a stacked dot plot or histogram, and a listing of statistics such as the five-number summary or the mean and standard deviation. When many data sets are being compared, box plots are a convenient visualisation if the representation is completely understood.



Note that the box plot for pulse rates after exercise shows the pulse rate of 146 as a possible outlier (.). This is because the distance of this data point above the upper quartile 42 (146-104) is more than 21 (1.5 × IQRs = 1.5 × (104 – 90) = 1.5 × 14 = 21).
The term 'parallel box-and-whisker plots' is commonly abbreviated to 'parallel boxplots'.

Figure 3. ACARA (2012) definition of Parallel box plots.

Box plot of number of blinks



If the inter-quartile distance is denoted by d, then the whiskers go out to the last data point inside the distance 1.5d from the edges of the box. Any data points outside this distance from the box are marked by *'s.

Figure 4. Box plot format recommended by MacGillivray (2011).[1]

[1] NB: In a separate document, Additional Information to Support the Glossary (ACARA, 2012) (available at www.australiancurriculum.edu.au/Mathematics/Additional-glossary-information) more detail is given about dealing with outliers under the definition of Box plot. The description, however, should include "at least" as shown in the following sentence: In constructing box plots, it is common to designate data values that lie a distance of at least 1.5 x IQR from either box end as possible outliers.

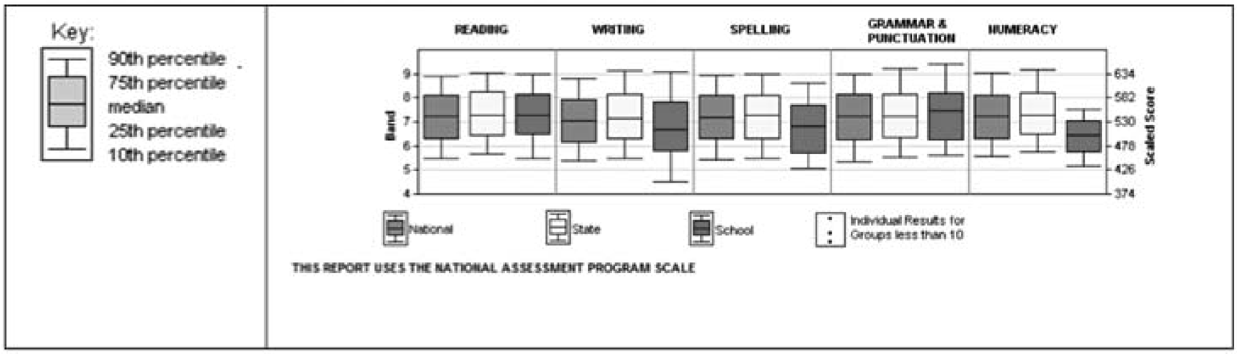


Figure 5. VCAA box plots for NAPLAN results.

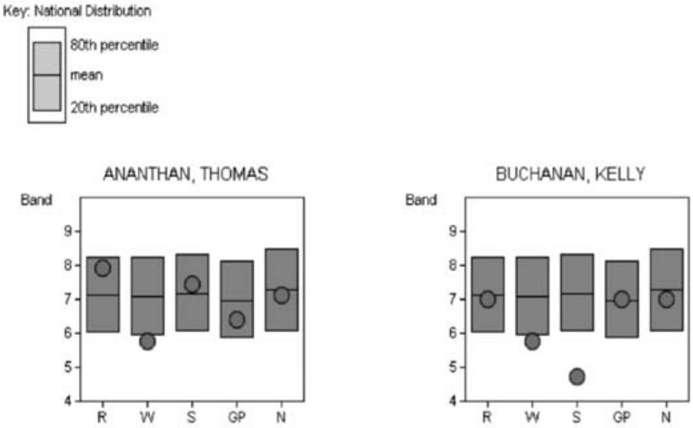


Figure 6. VCAA representations for NAPLAN scores for two imaginary students.

Student difficulties with box plots

As part of developing the pedagogical content knowledge necessary to teach about box plots, teachers need an appreciation of students as learners (Callingham & Watson, 2011) in relation to the concepts involved. By Year 10 it would be expected that students know the definitions and can distinguish between mean and median as measures of centre. These concepts are introduced in Year 7 of the curriculum and employed in various ways in Years 8 and 9. Research has shown, however, that the intuitions associated with the concepts often are not consolidated (Mokros & Russell, 1995; Watson & Moritz, 2000) and it is not natural for students to imagine using the measures for example as evidence to support the difference in two data sets (Watson & Moritz, 1999). As seen in the Glossary extract in Figure 2, percentiles, and hence percents, are important in the creation of box plots and the median is equal to the 50th percentile. Traditionally, of the three rational number representations—fractions, decimals, and percentages—percentages have received the least attention in the classroom. Research confirms that percentages cause considerable difficulty for students (Parker, 2004) and teachers need to be prepared to provide a review of percent linked to percentiles and creating box plots. Although not specifically mentioned in the curriculum the ogive, or cumulative relative frequency plot, can be a way of developing concrete understanding of percentiles (Siegel, 1988).

The most difficult aspect of straightforward box plots, as defined in Figure 1 based on the five-number summary, is the density representation in each of the four parts of the plot. Since each of the four quarters of the plot represent (approximately) 25% of the data, the linear length of the four segments represents the spread of the 25% rather than a larger or smaller number of data values. This is non-intuitive for many students who, for example, are used to a longer length representing "more" centimetres and a shorter length representing "fewer" centimetres on a ruler. The usual representation of a box plot without the accompanying data is a very useful shorthand for those who understand this inverse relationship: the smaller the associated length of the component of the plot, the *more closely packed* the same member of values are compared to the other components. This difficulty, reported by Bakker, Biehler, and Konold (2005), led them to suggest that box plots not be introduced to middle school students. Although one might expect that by Year 10 the density concept would be understood, Pierce and Chick (in press) found that many teachers in their study continued to experience difficulty as professionals interpreting system data.

Although for statisticians there is no need to *see* the data that create a box plot, for beginners it would appear useful to be able to do so. Three box plots created with the software *TinkerPlots* (Konold & Miller, 2011) are shown in Figure 7. In the top box plot, matching the main definition of box plot in the curriculum (Figure 1), 60 data values are shown under the plot, whereas in the middle plot the dividers also highlight the 25% of data in each quartile. In the bottom plot the data are hidden, giving the students a chance to imagine recreating the data, which can then be revealed again for reinforcement. The data in Figure 7 are symmetrically distributed and the lengths of each of the four components are roughly the same. Plots like this one are common but do not highlight the density issue.

The top plot in Figure 8 shows the box plot and the associated data set for Mark Taylor's batting scores for his first innings in his first 63 cricket test matches

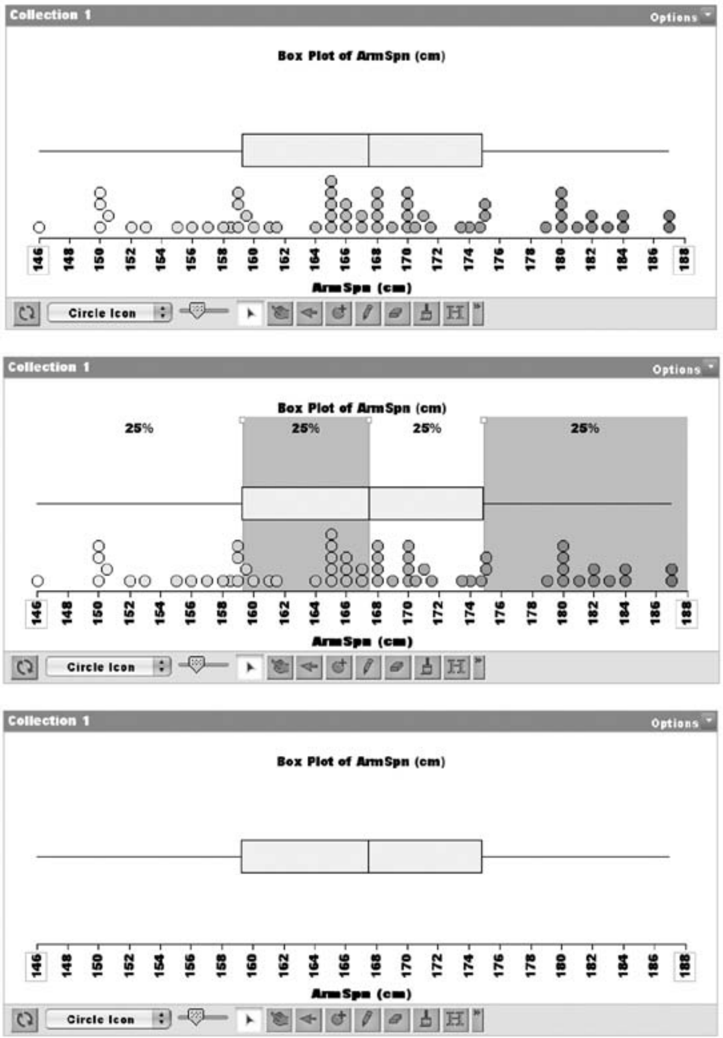


Figure 7. Box plot with data, box plot with data and dividers, and box plot without data.

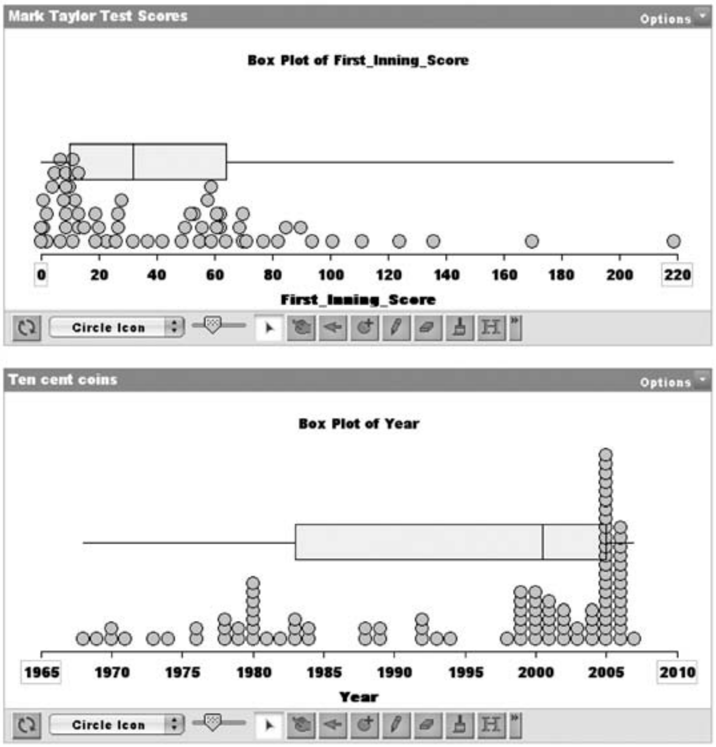


Figure 8. Box plots with skewed data sets.

(available at www.statsci.org/data/oz/taylor.txt). These data are skewed to the right with a very short whisker to the left and a very long whisker to the right. Without the data visible many students may judge by the length of the right whisker that Mark Taylor was a great batsman because *most* of his innings resulted in high scores. With the data visible, however, it can be seen that the data under the left whisker are very bunched up whereas the data under the right whisker are very spread out. Hence students should observe that the best quarter of Mark Taylor's innings resulted in scores more varied than all of the other three-quarters of his innings. The data in the second quartile are similarly more bunched up than the data in the third quartile but the difference is not as extreme. The plot at the bottom of Figure 8 shows the box plot and data set for the dates on 100 randomly collected Australian 10 cent coins. This distribution is skewed in the opposite direction to the cricket scores. Again the spread-out then bunched-up nature of the data is seen as the eye moves from left to right along the axis. The contexts of these two examples provide an excellent opportunity for teachers to discuss why skewness occurs in some data sets. Contrasting the shapes of the data distributions in Figures 7 and 8 allows some introductory discussion of normal distributions (e.g., Figure 7) and where they arise in contrast to asymmetric distributions. Being able to hold such a discussion using only the box plots without the actual data present is a sign of high level functioning and understanding of what box plots represent.

For students who struggle with box plots or for younger students there is a more basic representation, known as the hat plot, provided in *TinkerPlots*. As seen in Figure 9 in comparison to Figure 7, the median is not represented and the whiskers drop down to become the brim of the hat. The central 50% of the data are now clustered under the crown of the hat. Watson, Fitzallen, Wilson, and Creed (2008) showed that students in Years 5 to 7 could use hat plots to describe and compare data sets and develop language for the middle such as "clustered" and "squeezed up." Talking about the middle 50% of the data in the crown also reinforced links to percent, which had been taught to some of the students recently.

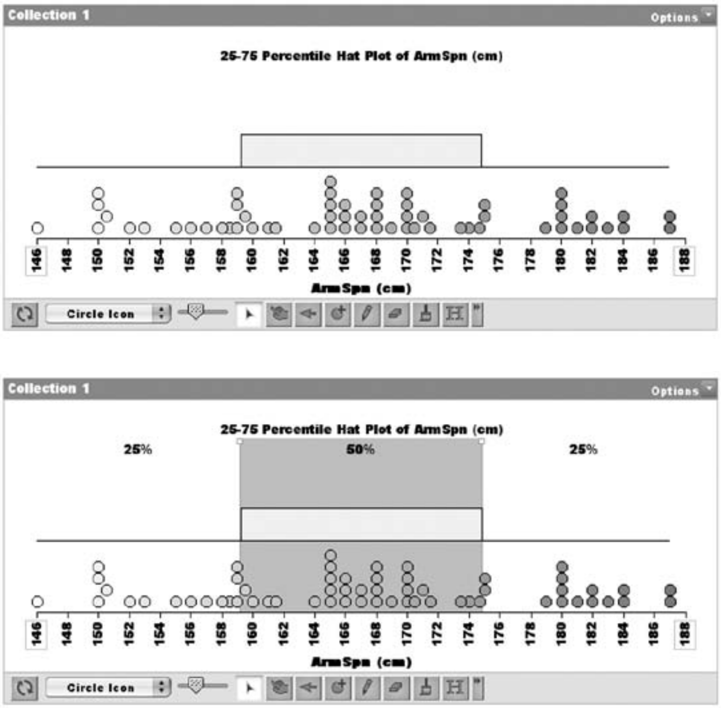


Figure 9. Hat plot with data and with dividers.

There are many aspects of learning about box plots that require careful planning and reinforcement by teachers. The final constant reminder is to look for and read a key if it is present.

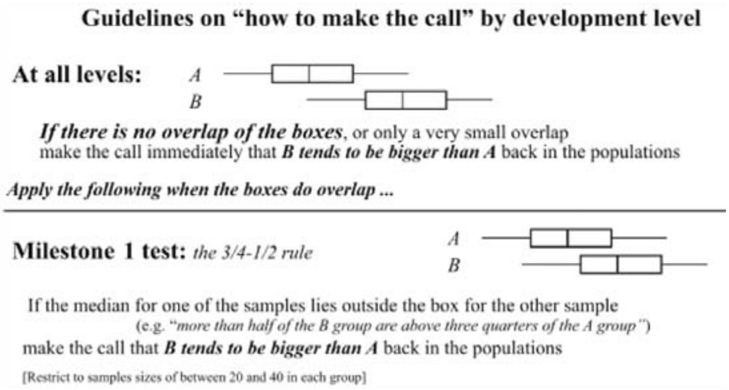
Table 1. Descriptors for Year 10 related to Box Plots (ACARA, 2012, p. 63)

Data representation and interpretation	Elaborations
Determine quartiles and interquartile range (ACMSP248)	finding the five-number summary (minimum and maximum values, median and upper and lower quartiles) and using its graphical representation, the box plot, as tools for both numerically and visually comparing the centre and spread of data sets
Construct and interpret box plots and use them to compare data sets (ACMSP249)	understanding that box plots are an efficient and common way of representing and summarising data and can facilitate comparisons between data sets using parallel box plots to compare data about the distribution of Aboriginal and Torres Strait Islander people by age with that of the Australian population as a whole
Compare shapes of box plots to corresponding histograms and dot plots (ACMSP250)	Investigating data in different ways to make comparisons and draw conclusions

Use of box plots

Table 1 contains the three descriptors and four associated elaborations related to box plots in Year 10 of the *Australian Curriculum: Mathematics*. The verbs "construct," "interpret," and "compare" are indicative of what teachers can plan to do with box plots in the classroom and the last two words are "draw conclusions." What is missing, however, is an indication of *how* box plots "can facilitate comparisons between data sets." There is no indication of criteria to be used in drawing conclusions when making comparisons.

Research has shown that when comparing two data sets, students often are initially reluctant to declare a likely difference if there is any overlap of the data values in a plot. Once over that hurdle, however, sometimes they will declare a very small, perhaps negligible, difference to be "real" (Watson, 2008). Wild, Pfannkuch, Regan, and Horton (2011) provide some guidance on making comparisons based on random samples of different sizes using box plots. Although some of their criteria may be considered too detailed for Year 10, several can be very useful. If there is no overlap for two box plots, including the whiskers, then it is safe to conclude there is a difference in populations underlying the two data sets. Wild et al. first base "making a call" about difference on the criteria in Figure 10. The $\frac{3}{4}$-$\frac{1}{2}$ rule is particularly useful.



Guidelines on "how to make the call" by development level

At all levels:

If there is no overlap of the boxes, or only a very small overlap make the call immediately that ***B tends to be bigger than A*** back in the populations

Apply the following when the boxes do overlap ...

Milestone 1 test: *the 3/4-1/2 rule*

If the median for one of the samples lies outside the box for the other sample (e.g. *"more than half of the B group are above three quarters of the A group"*) make the call that ***B tends to be bigger than A*** back in the populations

[Restrict to samples sizes of between 20 and 40 in each group]

Figure 10. Decision-making with box plots (Wild et al., 2011, Figure 9, p. 260).

The left of Figure 11 shows a plot of travel time to school for 30 randomly selected students (from the ABS Census@School site) who travel to school in three different ways. It is not possible to claim a difference in travel time for Australian students who walk or come in a car but because the box for bus is to the right of the box for car, it is possible to claim that travel time by bus tends to be longer than travel time by car for Australian students. The right of Figure 11 shows an example of the $\frac{3}{4}$-$\frac{1}{2}$ rule for dominant-hand

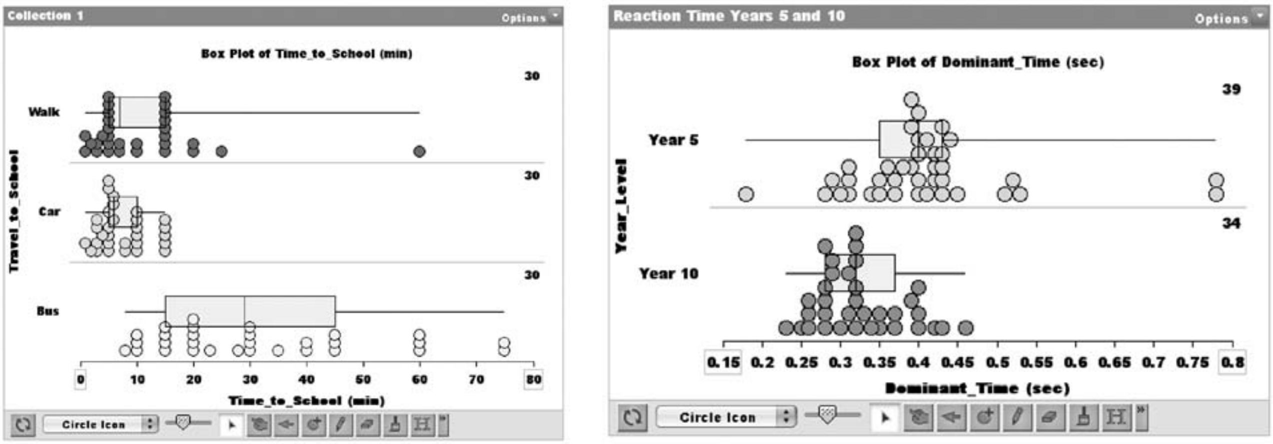
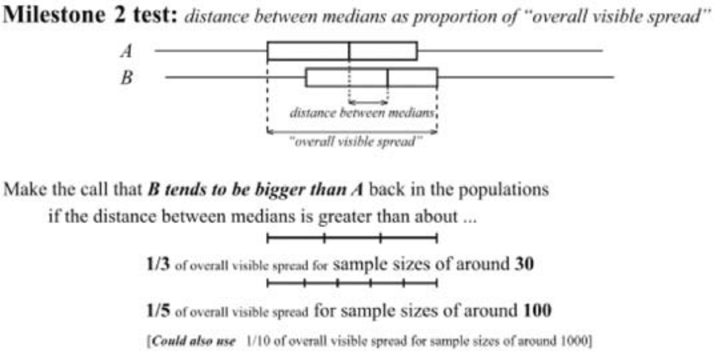


Figure 11. Box plots illustrating Wild et al.'s criteria for decision making.

reaction time data randomly collected for boys in Year 5 and Year 10 (from the ABS Census@School site). Since the median for Year 10 is outside the box for Year 5 (also true for the Year 5 median in relation to Year 10), in each case half of the data from one set are more extreme than three quarters of the data in the other set. It is hence possible to conclude that the dominant-hand reaction time for Australian Year 10 boys tends to be faster than the dominant-hand reaction time for Australian Year 5 boys. These visual criteria provide concrete cues for students in drawing conclusions, as suggested should be done in the elaboration in Table 1. They could even be used with care for comparing data sets like those shown in Figure 5.

The next stage in decision-making for Wild et al. (2011) is more complex and is shown in Figure 12. It starts to become more procedural and less intuitive and might be suggested for more advanced students to use in project work. These criteria for decision-making are preliminary to formal inference, which students may meet in Years 11 and 12 or university. The suggestion of criteria to make decisions based on box plots however is important in introducing students to the use of evidence from samples to make informal inferences about populations, acknowledging uncertainty (Makar & Rubin, 2009).


Milestone 2 test: *distance between medians as proportion of "overall visible spread"*
A
B
distance between medians
"overall visible spread"

Make the call that ***B tends to be bigger than A*** back in the populations if the distance between medians is greater than about ...

1/3 of overall visible spread for sample sizes of around **30**

1/5 of overall visible spread for sample sizes of around **100**

[***Could also use*** 1/10 of overall visible spread for sample sizes of around 1000]

Figure 12. Advanced decision-making with box plots (Wild et al., 2011, Figure 9, p. 260).

Summary

Although box plots in various forms have been taught in some jurisdictions in the past, the *Australian Curriculum: Mathematics* (ACARA, 2012) provides the opportunity to consolidate what is taught across the country. Given the variations in the formats for box plots that students are likely to meet in other curriculum areas and outside of school, flexibility is important. Box plots are different from other mathematical objects that can be more precisely defined. If ACARA develops an assessment program employing box plots it will be necessary to make clear that the version described in the glossary is to be assessed and tasks will need to be devised based on data sets where "approximately" does not make judging responses difficult. One concern arises when observing the lack of criteria for drawing conclusions in Table 1. Will assessment items be restricted to low level procedural tasks? This would be a shame given the potential of box plots to lead students to a high level of statistical thinking.

Acknowledgements

The author thanks colleagues on the ARC Linkage project LP100100388, Ian Gordon, Robyn Pierce, and Helen Chick, for useful discussions on the issues addressed in this article.

References

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2012). *The Australian Curriculum: Mathematics, Version 3.0, 23 January 2012.* Sydney, NSW: ACARA.

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 Statistical Education (IASE) Roundtable, Lund, Sweden, 28 June – 3 July 2004* (pp. 163–173). Voorburg, The Netherlands: International Statistical Institute.

Callingham, R. & Watson, J. (2011). Measuring levels of statistical pedagogical content knowledge. In C. Batanero, G. Burrill & C. Reading (Eds), *Teaching statistics in school mathematics — Challenges for teaching and teacher education: A joint ICMI/IASE study* (pp. 283–293). Dordrecht, The Netherlands: Springer.

Konold, C. & Miller, C. D. (2011). *TinkerPlots: Dynamic data exploration* [computer software, Version 2.0]. Emeryville, CA: Key Curriculum Press.

MacGillivrary, H. (2011). *Data investigation and interpretation: A guide for teachers – Year 10. Statistics and Probability: Module 8.* Melbourne: The Improving Mathematics Education in Schools Project. Available at http://www.amsi.org.au/teacher_modules/Data_investigation_year_10.html

Makar, K. & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82–105.

Mokros, J. & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, *26*, 20–39.

Parker, M. (2004). Reasoning and working proportionally with percent. *Mathematics Teaching in the Middle School, 9,* 326–330.

Pierce, R. & Chick, H. (in press). Workplace statistical literacy for teachers: Interpreting box plots. *Mathematics Education Research Journal.*

Siegel, A. (1988). *Statistics and data analysis: An introduction.* New York: Wiley.

Tukey, J. W. (1970). *Exploratory data analysis.* [Preliminary ed.] Reading, MA: Addison-Wesley.

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

Victorian Curriculum and Assessment Authority. (2010). *Reporting guide: 2010 National Assessment Program Literacy and Numeracy.* East Melbourne: Author.

Wall, J. J. & Benson, C. C. (2009). So many graphs, so little time. *Mathematics Teaching in the Middle School*, *15*, 82–91.

Watson, J. M. (2008). Exploring beginning inference with novice grade 7 students. *Statistics Education Research Journal*, *7*(2), 59–82.

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*, 4–10.

Watson, J. M. & Moritz, J. B. (1999). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, *37*, 145–168.

Watson, J. M. & Moritz, J. B. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning*, *2*(1&2), 11–50.

Wickham, H. & Stryjewski, L. (2011). *40 years of box plots.* Manuscript submitted for publication. Available at: http://vita.had.co.nz/papers/boxplots.html

Wild, C. J., Pfannkuch, M., Regan, M. & Horton, N. J. (2011). Towards more accessible conceptions of statistical inference (with Discussion). *Journal of the Royal Statistical Society A, 174*(2), 247–295.