REFEREED PAPER

COORDINATING BETWEEN HISTOGRAMS AND BOX PLOTS

LEM Stephanie, ,
Centre for Instructional Psychology and Technology, Katholieke Universiteit Leuven
stephanie.lem@ped.kuleuven.be
ONGHENA Patrick
Methodology of Educational Sciences Research Group, Katholieke Universiteit Leuven
VERSCHAFFEL Lieven
Centre for Instructional Psychology and Technology, Katholieke Universiteit Leuven
VAN DOOREN Wim
Centre for Instructional Psychology and Technology, Katholieke Universiteit Leuven

ABSTRACT

Representational fluency consists of several aspects, like the efficiency with which one uses a particular representation and the efficiency with which one is able to coordinate between different representations. In this study we focused on this last element, more specifically on coordinating between histograms and box plots. Participants were 167 first year university students. They were asked to match box plots and histograms of the same distributions and to explain their matches. We found that students had one major difficulty when interpreting box plots: They tended to interpret the area of box plots incorrectly by assuming that a larger area represented more observations than a smaller area. In both items this led to incorrect matches and to incorrect explanations of these matches. Furthermore, students displayed several other misinterpretations.

INTRODUCTION

The correct use of graphs and other external representations is an important goal of statistics education. This representational competence consists of two main elements: representational fluency and representational flexibility (Heinze, Star, & Verschaffel, 2009). The latter refers to being able to choose the optimal representation for a certain task. The former concerns several aspects, such as the efficiency with which one uses a particular representation and with which one is able to coordinate between different representations. This study focused on this last skill: coordinating between different representations. It has been shown that this coordination between representations is a skill students sometimes lack, causing confusion and lower achievement on mathematics tasks (Leinhardt, Zaslavsky, & Stein, 1990). However, we think that also a lack of efficiency (especially the misinterpretation of representations) can hinder coordination between representations. In this paper, we will discuss how the (lack of) efficiency with which students use histograms and box plots, operationalised as the accuracy with which they interpret the representations, prevents them from correctly coordinating between representations.

We chose to work with representations of data distributions, as (data) distributions are seen as an important concept students have to master. There are many representations available to present data distributions. Recent studies have shown that students have great difficulty interpreting these representations, especially histograms and box plots.

With regard to histograms, delMas, Garfield, and Ooms (2005) found, in a large-scale assessment of the errors high school and college students made in the interpretation of histograms, three errors students frequently displayed. First, students often read the horizontal axis of the histogram as a time scale, leading them to conclusions about evolutions of the data, which are not displayed. Second, students tended to confuse histograms with bar graphs, leading them to the interpretation of each rectangle as representing one observation of which the vertical axis reflects the value of the observed variable. Finally, students not always recognized the *groupedness* of grouped histograms, leading to erroneous responses when asked to read off frequencies of specific values. This last misinterpretation was also observed in university students by Lem, Onghena, Verschaffel, and Van Dooren (2011), both with respect to histograms and box plots. Watson and Moritz (1998) found that many students in grade 3 to 9 compared the frequency or height of the modes of two histograms when trying to compare the means of these histograms, which was also found in university students by Lem, Onghena, Verschaffel, and Van Dooren (2010) and Lem et al. (2011). Meletiou and Lee (2002) found that students in an introductory statistics course, when asked to compare the variation of two histograms, tended to interpret the height differences between the different rectangles of the histograms as representing the variation. This same misinterpretation was found by Cooper and Shore (2008) and by Lem et al. (2010).

Regarding box plots, Bakker, Biehler, and Konold (2004) found that students in seventh and eighth grade of secondary school tended to interpret the area in box plots as reflecting the frequency or proportion of observations. Variations of the box plot exist in which the area of the box is proportional to another variable, but in the standard box plot the area only reflects density. Bakker et al. (2004) observed that this reflection of density instead of frequency or proportion makes box plots very different from histograms and bar graphs. Lem et al. (2010; 2011) found the same misinterpretation in first year university students.

In this study we investigated how the interpretation difficulties of histograms and box plots as reported in the literature may prevent students from correctly coordinating between histograms and box plots.

METHOD

Participants were 167 first year university students (158 female, 9 male) of Educational Sciences at the Katholieke Universiteit Leuven, Belgium. In return for their participation they received course credit. Prior to their participation, all participants had taken the same introductory statistics course, covering descriptive statistics, graphical representations, and distribution.

The here presented items were part of a larger paper-and-pencil test on external representations for data distributions. All items in the test were constructed based on a pilot study. In both items, students were asked to match box plots and histograms representing the same data distribution. In the first item, students were asked to match a histogram of a skewed data distribution to its corresponding box plot, while in the second item, students were asked to match histograms of symmetric data distributions with different variation to their corresponding box plots. Besides giving the matches, we also asked the students to explain how they found the matching representations. The exact items will be presented with their respective results in the results section.

The analysis of students' explanations of their reasoning was done by the first author. Based on the misinterpretations students displayed in the pilot study, an initial coding scheme was constructed. During the coding, some other misinterpretations came up, which were then included in the coding scheme. For each explanation, we coded which misinterpretation(s) was present. Only the explanations for the incorrect matches were coded.

RESULTS

ITEM 1

In this first item, students were asked to link one of two box plots to the shown histogram (Figure 1). Both box plots were exactly each other's opposites, box plot A being the box plot representing the same data as the histogram. No scale was presented on the axes in order to prevent students from trying to compare some measure of center. Because of the large area in the left side of box plot B, we expected that students might think this represented the high peak on the left side of the histogram. This reasoning, larger area stands for more observations, holds for histograms, but not for box plots. This because box plots, as opposed to histograms, present densities and not frequency or proportion (Bakker et al., 2004).

One of these box plots represents the same data as the histogram. Which one is it? How do you know?

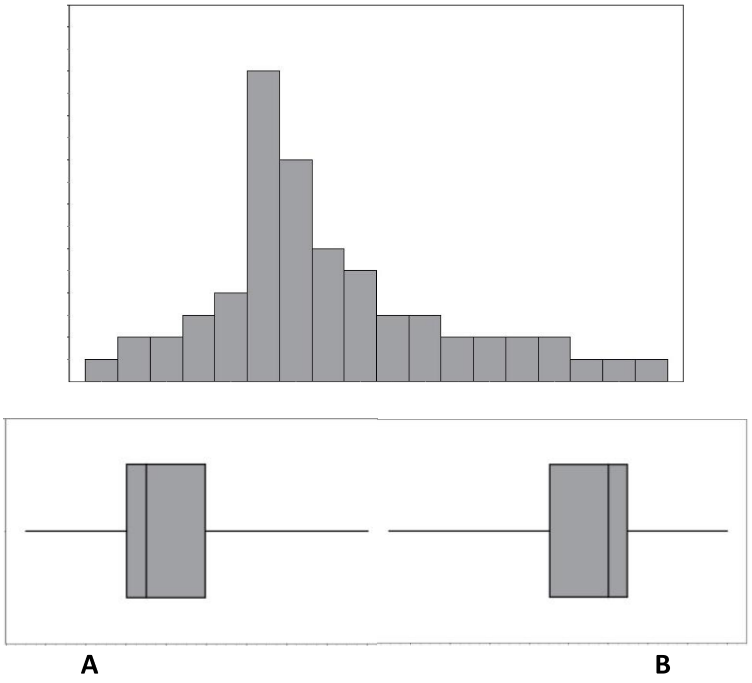

Figure 1. Item 1

This item was solved quite well, as 81% of the students provided the correct match. Of the 19% students not providing the correct match, 48% did not provide an explanation for their response, indicating that they guessed, or provided a very unclear explanation that we were not able to understand and categorize. Furthermore, 31% of these students explained, as we expected, how there were more observations represented in the left part of the box in box plot B, with the largest area, than in the right part. Finally, 21% of the students indicated that the median should be more to the right of the distribution, indicating they were not able to estimate the value of the median correctly, either due to a misinterpretation of the histogram or due to a misunderstanding of the concept of the median.

Students' explanations for their correct matches suggest they often looked at the right whisker of box plot A, comparing it with the long tail of the histogram. In the second item, however, the tail could not cue the students towards the correct match in this way.

ITEM 2

In item 2, we asked students to match two symmetrical box plots with different variation to two histograms. Box plot A represents the same data as histogram C, a unimodal symmetrical distribution. Box plot B and histogram D also represent the same data, which forms a bimodal symmetrical distribution. The variation in box plot B and histogram D is higher than the variation in box plot A and histogram C. On first sight, it might look like the short whiskers in box plot B, represent the low bars on the extremes of histogram C, as a larger area could be associated with a larger number of observations. The small rectangles in box plot A, might look like representing the lower bars in the middle of histogram D. However, because box plots represent density and not frequency or proportion, this reasoning is not valid.

Match the box plots with their corresponding histogram. Explain how you made the matches.

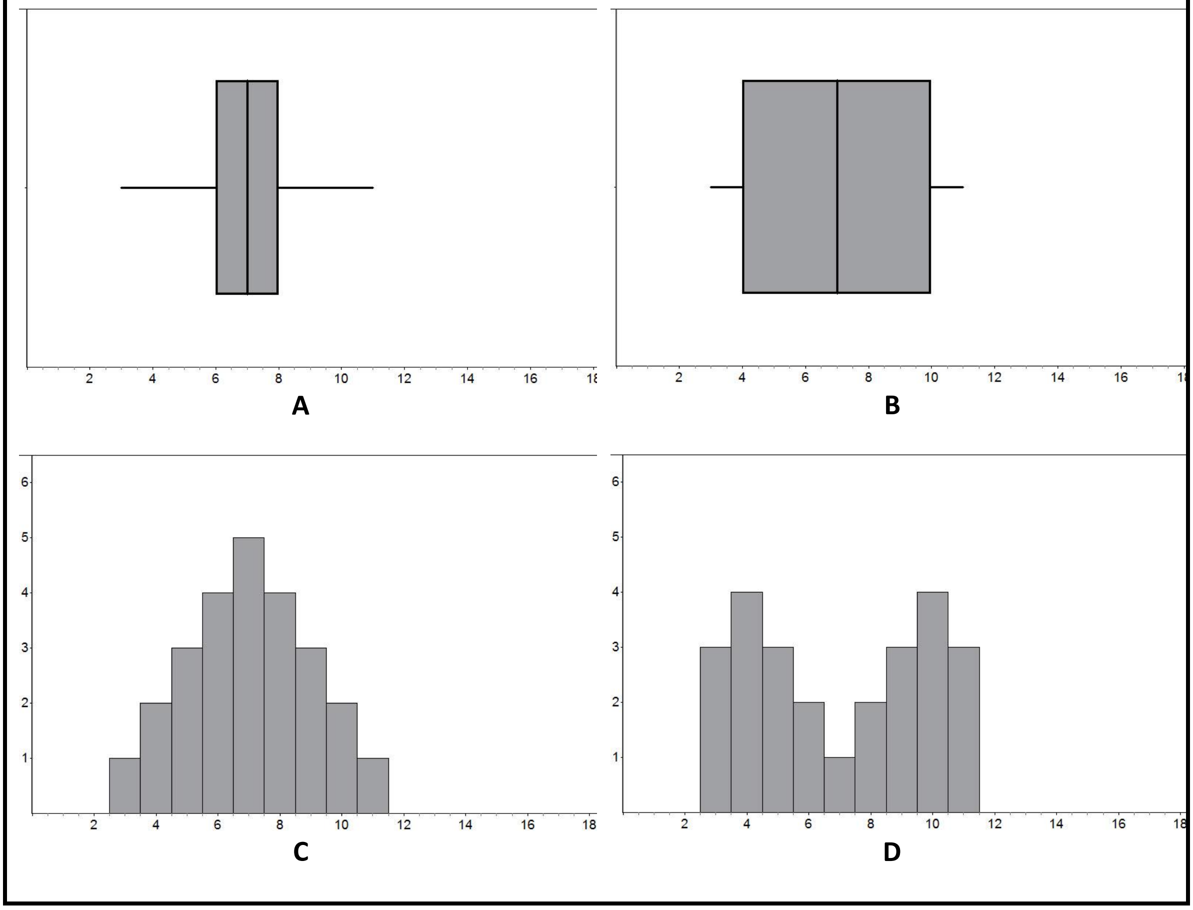

*Figure 2***.** Item 2

As only 58% of the students provided the correct solution, this item was more difficult than Item 1. Of the 42% students not providing the correct matches, 52% stated that they guessed, did not provide any explanation, or gave an unclear explanation. Furthermore, 46% of these students stated in their explanation that the larger area in the box plot represented a larger amount of observations in that interval. Finally, 1 student explicitly stated that Q1 was the lowest observed value and Q3 the highest observed value, neglecting about half of the data.

DISCUSSION

In this study we investigated which misinterpretations prevented students from making the correct matches between histograms and box plots. By asking students to match histograms and box plots representing the same data distributions and having them explain their answers, we were able to observe how students' misinterpretations prevented them from matching histograms and box plots representing the same data distribution.

First, students interpreted the area of the box plot as representing frequency or proportion of observations: a larger area reflects more observations than a smaller area. This result is in accordance with the observation of Bakker et al. (2004) and the results of Lem et al. (2010, 2011). This misinterpretation led students to make incorrect matches as they incorrectly assumed specific parts of the box plot to represent many observations, matching the box plot to a histogram with high bars in that specific interval. Second, one student explicitly stated that Q1 was the lowest observed value represented in the box plot. It is not unlikely more students thought this, but our items were not constructed to test for this misinterpretation. A previous study (Lem, et al., 2011) even showed that almost one fifth of the students misinterpreted the box plot in this way. This misinterpretation leads to confusion as the minimum of the box plot and the histograms seem to be different, leading students possibly to incorrect matches or guessing. Finally, some students were not able to estimate the median of a distribution using the histogram. This might either be due to a misinterpretation of the histogram itself, or due to a misunderstanding of the concept of median. Again, this lead to incorrect matches as students often tried to find the matching box plot, in which they could read off the median, by estimating the median in the histogram.

Various interesting research questions are still to be answered. First, it would be interesting to gain a more in-depth insight into students' reasoning, for instance by interviewing students while solving problems and asking them to think aloud. Second, it would be interesting to see which other misinterpretations occur in students and whether students at different educational levels and fields of study show these misinterpretations to the same extent. Similar studies could be done in other areas of statistics education, like probability and hypothesis testing, but also in various other domains, like mathematics and science education. Finally, the role of representations seems somewhat underestimated in statistics education. When studying students' understanding of a specific statistical concept, this understanding must usually be studied by using external representations. Hence, it is often unclear whether students misinterpret the representation or misunderstand the underlying statistical concept, which is something that must be clarified in order to get a better understanding of students' reasoning.

With regard to education, we can say that, before students are confronted with both histograms and box plots to solve a certain task, it is important that they are able to correctly interpret each of these representations separately. On the other hand, however, the use of both representations together might be a fruitful way of letting students understand the merits and pitfalls of both representations, giving them a chance to construct a more coherent understanding of both representations. Furthermore, an insufficient understanding of histograms and box plots might hinder the learning of other subjects, such as the incorrect interpretation of one of these representations presenting a certain phenomenon, might cause a misunderstanding of this phenomenon.

Concluding, we can state that the interpretation of histograms and box plots is not as straightforward as is often assumed. Histograms and box plots are not only used in all levels of statistics education, students are also assumed to understand them in other courses, just like researchers and other professionals need to be able to use and interpret them as part of their job. Even the wider public is increasingly frequently confronted with these representations. The American Psychological Association even recommended the use of box plots in scientific publications (Wilkinson & The task force on statistical inference, 1999). Our study shows that a better understanding of histograms and box plots is needed.

ACKNOWLEDGMENTS

Stephanie Lem holds a PhD fellowship of the Research Foundation - Flanders (FWO).This research was partially supported by Grant GOA 2006/01 "Developing adaptive expertise in mathematics education" from the Research Fund K.U.Leuven, Belgium.

REFERENCES

Bakker, A., Biehler, R., & Konold, C. (2004). Should young students learn about box plots? In G. Burrill (Red.), *Curricular development in statistics education: International Association for Statistical Education* (pp. 163–173). Presented at Curricular development in statistics education: International Association for Statistical Education, Lund, Sweden.

Cooper, L. L., & Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education, 16*(2).

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In *Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy. University of Auckland*.

Heinze, A., Star, J., & Verschaffel, L. (2009). Flexible and adaptive use of strategies and representations in mathematics education. *ZDM Mathematics Education*, *41,* 535-540.

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research*, *60*(1), 1–64.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2010). *External representations for data distributions: in search of cognitive fit.* Manuscript submitted for publication.

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2011). *Why a graph is (sometimes) not worth ten thousand words: A study on the misinterpretation of histograms and box plots.* Manuscript submitted for publication.

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. *In Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa*.

Watson, J. M., & Moritz, J. B. (1998). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145–168.

Wilkinson, L., & The task force on statistical inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, *54*, 594-604.