Representation Use in Statistical Problem Solving

Stephanie Lem
Centre for Instructional Psychology and Technology, K.U.Leuven
Andreas Vesaliusstraat 2 – PO box 3770 – 3000 Leuven, Belgium
Stephanie.Lem@ped.kuleuven.be

Wim Van Dooren
Centre for Instructional Psychology and Technology, K.U.Leuven
Andreas Vesaliusstraat 2 – PO box 3770 – 3000 Leuven, Belgium
Wim.VanDooren@ped.kuleuven.be

Patrick Onghena
Centre for Methodology of Educational Research, K.U.Leuven
Andreas Vesaliusstraat 2 – PO box 3762 – 3000 Leuven, Belgium
Patrick.Onghena@ped.kuleuven.be

Lieven Verschaffel
Centre for Instructional Psychology and Technology, K.U.Leuven
Andreas Vesaliusstraat 2 – PO box 3770 – 3000 Leuven, Belgium
Lieven.Verschaffel@ped.kuleuven.be

Abstract. A lot of attention has been paid to the role of external representations in learning and problem solving, especially within mathematics. Much less research has been conducted, however, on the role of external representations in statistical problem solving. Research has focused on students' difficulties when reasoning about distributions, but the role of external representations has not yet been clarified. The study presented here focused on university students' fluent use of external representations for comparing the distributions of two data sets. By giving the same items with different representations (dot plot, histogram, box plot, and descriptive statistics) to 167 first year university students, we were able to confirm the hypothesis that the external representation used in distributional problem solving affects the reasoning and the achievement of students on a task. We also found a number of misinterpretations or misconceptions regarding these representations.

Keywords: external representations; representational fluency; problem solving; statistics; data distributions

Introduction

The mathematics education literature (e.g., Goldin, 2002) provides empirical evidence on the role of multiple representations in learning and problem solving in mathematics. In the statistics education literature much less attention has been paid to external representations and empirical research is scarce. The aim of the study (partly) presented here, was to fill this gap by studying students' problem solving with various external representations of data distributions.

Many statistical concepts can be used to study the role of external representations. The concept of "distribution", however, is especially suitable for this purpose: It is almost impossible to think about a distribution without using a representation, like a histogram or a box plot. Additionally, the concept of distribution is crucial in the statistics curriculum, and reasoning about this concept was shown to be difficult for students (e.g., Garfield & Ben-Zvi, 2008). Only some anecdotic evidence for difficulties and misconceptions in interpreting box plots exists (e.g., Bakker, Biehler, & Konold, 2004). We systematically investigated the effect of given representations on the accuracy of reasoning, the interaction with the kind of task, and the misconceptions students show in using representations.

Method

A total of 167 first year university students of educational sciences participated in return for course credit. They had followed one introductory statistics course (in which, among other topics, attention was paid to descriptive statistics, graphical representations, and distributions) prior to their participation. This paper focuses on the results of a part of a larger paper and pencil test. More specifically, we will focus on three items in which students were asked to compare two brands of batteries of which the distribution of the life spans was shown. Comparisons had to be made with respect to a given characteristic of the distributions: the mean, variation, or skewness. The representation that accompanied the question was randomly varied between students: one fourth of the students compared the mean of two distributions of which the dot plot was shown, one fourth received grouped histograms, one fourth box plots, and one fourth received tables with descriptive statistics (mean, median, mode, standard deviation, quartiles, minimum, maximum, and skewness). The same happened for comparison of the variation and skewness. These four representations were chosen because of their different levels of abstractness (Roth & Bowen, 1994). For each item the student was asked to give the correct answer (multiple choice) and to explain how (s)he found that answer.

Results

Table 1 presents the performance per item and per representation. Using the descriptives representation as a point of comparison (in this representation the answer could, in principle, be directly read off), we see a couple of deviating accuracy rates. We will discuss these below.

Table 1: Percentages of correct answers per item and per representation.

	Dot plot	Histogram[1]	Box plot	Descriptives
Mean	58.5	35,6	63.4	82.5
Variation	71.8	58.5	73.8	72.1
Skewness	43.,2	35.0	17.1	67.5

In comparing the mean of two distributions (see Figure 1), a lot of mistakes were made in all three graphical representations, compared to students who received the descriptive statistics. Especially the histogram led to a lot of errors. The explanations of the students reveal that in all representations some students used the range (e.g., “a larger maximum means a higher mean”) and variation (e.g., “a smaller variation means a higher mean”) of a distribution to compare the means. But for the histogram, an additional problem occurred: Students tended to think that the distribution in which the bar containing the mean was higher, also had a higher mean. In the dot plot the same mistake occurred, but less often than with the histogram (maybe because the differences in bar height of a grouped histogram are more pronounced than the differences between the heights of the stacked dots).

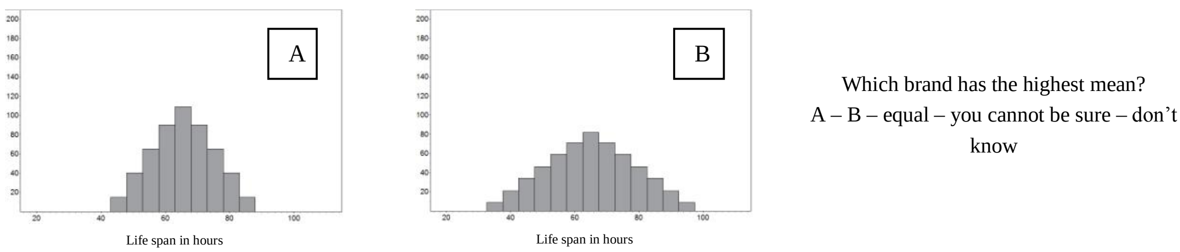

Figure 1. Example item. Students are asked to compare the mean life span of two brands of batteries.

[1] As we used grouped histograms, the answer to the ‘mean’-item was “you cannot be sure of the answer”. Because only two students correctly interpreted the bars and the class width, we coded ‘equal’ also as correct. This problem occurred also in many of the items not presented here and shows another problem with the interpretation of histograms than the ones we discuss in this paper.

In the variation item, students were asked to compare the variation of two symmetrical distributions with the same mean but a different range. Here only the histogram seems to lead to more errors than just providing the descriptive statistics. With all representations, a lot of students did not give any answer, which suggests that the notion of "variation" is difficult for them. With the histogram representation, a more specific error occurred: Four students (out of the 16 students giving an incorrect answer) looked at the "jumps" (differences in frequencies between neighboring bars) as an indicator of variation.

In the skewness item, participants were asked which battery brand (one distributed symmetrically, the other skewed) is best when you want a large chance of getting a life span longer than the mean (both distributions had the same mean). Again all three graphical representations caused a lot of difficulty for students, but the box plot was most troublesome. The main cause of errors using the box plot was that participants looked at the area of a part of the box as an indicator of the number of observations within this range (the larger the area the more observations), while each of the four parts of a box plot represents 25% of the observations. A problem that occurred in all four representations is related to the interpretation of the question as such, i.e. assuming that a higher maximum or a larger range would automatically lead to a larger chance of having a battery with a higher than average life span. Histograms and dot plots furthermore had another specific problem: Students tended to look at absolute numbers instead of relative numbers or chances. The more abstract box plot did not induce this misinterpretation.

Conclusion

The accuracy rates on the different item-representation combinations show that the representation used to compare two distributions affects the reasoning and the chance to answer correctly. It is thus important to consider the representations offered when assessing distributional problem solving. These results could also have implications for the teaching and learning of students. The explanations the participants provided for their answers revealed various misconceptions regarding the representations. Not only box plots are difficult to interpret for students (Bakker et al., 2005); other representations also evoke misconceptions in students, even though these representations look more intuitive and less abstract than box plots. More research is needed on the nature and the occurrence of these misconceptions.

References

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 2004 Roundtable* (pp. 163-173). Voorburg: International Statistical Institute.

Garfield, J. B., & Ben-Zvi, D. (2008). Learning to reason about distribution. In J. B. Garfield & D. Ben-Zvi. (Eds.), *Developing students' statistical reasoning: Connecting research and teaching practice* (pp. 165-186). Dordrecht: Springer.

Goldin, G. A. (2002). Representation in mathematical learning and problem solving. In L. D. English (Ed.), *Handbook of international research in mathematics education* (pp. 197-218). Mahwah, NJ: Erlbaum.

Roth, W., & Bowen, G. M. (1994). Mathematization of experience in a grade 8 open-inquiry environment: An introduction to the representational practices of science. *Journal of Research in Science Teaching*, *31*(3), 293-318.