Students' reasoning about sampling distributions before and after the Sampling Distribution Activity.

Vanhoof, Stijn
Department of Educational Sciences
Andreas Vesaliusstraat 2
3000 Leuven - Belgium
E-mail: stijn.vanhoof@ped.kuleuven.be

Castro Sotos, Ana Elisa
Department of Educational Sciences
Andreas Vesaliusstraat 2
3000 Leuven - Belgium
E-mail: anaelisa.castrosotos@ped.kuleuven.be

Onghena, Patrick
Department of Educational Sciences
Andreas Vesaliusstraat 2
3000 Leuven - Belgium
E-mail: patrick.onghena@ped.kuleuven.be

Verschaffel, Lieven
Department of Educational Sciences
Andreas Vesaliusstraat 2
3000 Leuven - Belgium
E-mail: lieven.verschaffel@ped.kuleuven.be

1. Introduction

Although the development of the idea of a sampling distribution is a critical step in developing the theory of statistical inference, students often have misunderstandings when learning about sampling distributions (Chance, delMas, & Garfield, 2004). One example of a widespread and important misunderstanding is *neglecting the effect of sample size on sampling variability.* For instance, in contrast to what probability theory would predict, most subjects in a study of Tversky and Kahneman (1974) judged the event of obtaining more than 60 percent boys to be the same in a small hospital with 15 and in a large hospital with 45 births a year. Another example of the problems students encounter is the *misunderstanding of the Central Limit Theorem*. For instance, in contrast to what this theorem states, many students in a study of Chance et al. (2004) believed that sampling distributions should look more like the population as the sample size increases.

To overcome these difficulties, the use of computer simulations for teaching statistics in general, and for teaching sampling distributions in particular, has been recommended by many researchers (for a review, see Mills, 2002). To investigate the possible advantages of computer simulation methods, Chance et al. (2004) and delMas, Garfield, and Chance (2004) have developed the so-called *Sampling Distribution Activity* (SDA), an instructional activity that includes *Sampling Sim*, simulation software specifically designed for teaching sampling distributions. Besides this simulation software, the activity includes several assessment tasks and a pre- and posttest to assess students' prerequisite knowledge and understanding of sampling distributions. Most items of the pre- and posttest are closely related to the assessment tasks in the activity. The present large-scale investigates the effect of the SDA on students' understanding of sampling distributions.

2. Research goals

Following the research tradition of Chance et al. (2004) and Delmas et al. (2004), the main goal of our study is to gain more insight into the effectiveness of the Sampling Distribution Activity for teaching sampling distributions. More specifically, the aim is to document students' understanding of sampling distributions before and after the Sampling Distribution Activity.

3. Method

Participants

Participants were 221 students of Educational Sciences or Speech Pathology at the Katholieke Universiteit Leuven in Belgium. The typical curriculum of these students covers five years. Students have to follow three statistics courses, one in each of the first three years. The second statistics course, which uses the introductory textbook of Moore and McCabe (2006), covers some methodology, probability, sampling distributions and an introduction in statistical inference. The mathematical background required to follow these statistics courses is limited.

Sampling Distribution Activity

In the second year statistics course, the *Sampling Distribution Activity* was used for teaching sampling distributions. Three modifications were made to the original activity to better fit our specific context. First, in contrast to previous studies, we used the activity as the very first introduction to the concept of sampling distributions. Second, where previous studies focus on the sampling distribution of the mean, we also included sampling distributions of other statistics like the proportion and the standard deviation. Third, a blackboard scheme was used to facilitate the integration of different ways to visualize the process of sampling distributions.

The complete activity took place in two two-hour sessions. The first session was organized in groups of about 40 students. It started with an interactive introduction to sampling distributions of proportions and means for the whole group, followed by small group exercises. More than half of the session was devoted to the use of *Sampling Sim* and to the graphical exercises of the *Sampling Distribution Activity* (see further). The teacher showed *Sampling Sim* and the exercises on one big screen. The teacher did not provide answers, but guided the students in finding answers to all questions. The second session took place in one large group and included an interactive lecture about more formal derivations of characteristics of sampling distributions followed by some exercises.

Instruments and data analysis

Most of the material used to assess students' understanding of sampling distributions originates from the pre- and posttest included in the SDA. First, there is a multiple-choice non-graphical contextual '*geology item*', where students have to assess the accuracy of the average of 5 versus 20 weightings to determine a rock's weight. It aims at measuring students' intuitive understanding of the impact of sample size on sampling variability. Second, there is a graphical item where students have to distinguish a sample and a sampling distribution. Students have to indicate whether there is a difference between elements in two figures (one sample and one sampling distribution) and, if they think there is, they have to explain this difference. Third, there are two graphical items that aim at measuring students' visual understanding of sampling distributions. For these items, students have to choose the best histogram of a sampling distribution, given a specific graph of the population and a sample size and they have to judge characteristics such as variability and shape. Both items include seven sub-questions. So, for these two items together students can have a score ranging from zero to fourteen. Since these last two graphical items require at least some previous experience with sampling distributions, they were only assessed in the posttest.

In addition, participants completed the *Statistical Reasoning Assessment* (SRA; Garfield, 2003). One item of this questionnaire is particularly of interest for our study, namely the above-mentioned *'hospital item'* from Tversky and Kahneman (1974). For this item, administrations at three different points in time are available: at the beginning of students' first year, and before and after the SDA.

In summary, pre- and posttest data (before and after the SDA in the second year) are available for two non-graphical items (*'hospital item'* and *'geology item*) and for one graphical item. For the 'hospital item' also data at the beginning of students' first year are available. Only posttest data are available for two other graphical items. Because not all students are able to follow all teaching sessions, not all 221 students participated at each administration. Descriptive data are presented for all available data. The response rate for the '*hospital item*' was 91% (n = 202) for the first, 72% (n = 160) for the second, and 93% (n = 205) for the third administration. The response rate for the other items (items of the SDA pre- and posttest) was 86% (n = 191) for the pretest and 81% (n = 179) for the posttest.

Randomization tests (Edgington & Onghena, 2007), more specifically exact randomization test versions of Cochran Q test and McNemar change test, are used to analyze data for students that participated at all administrations. For the hospital item this is the case for 69% (n = 152) of the students, for the 'geology item' for 73% (n = 162), and for the first graphical item for 65% (n = 143). The results are similar if we consider all available data or data for students that participated at all measurement moments.

4. Results

Non-graphical items

The results for the *'hospital item'* show that many students have difficulties to realize the impact of sample size on sampling variability properly. For all three administrations, only between 30% and 40% of the students judged the probability of obtaining a larger proportion of female births to be larger for a small hospital than for a large hospital. There was no statistically significant difference between the percentage of correct answers for the different observation moments (33.67% at the beginning of the first year, 31.88% before the SDA, and 37.07% after the SDA), Q(2, N = 152) = 3.6438, p = .1609.

For the '*geology item'*, the number of correct answers even decreases from pretest to posttest. In the pretest 88.48% of the students correctly indicate that the average of 20 weightings would be more accurate than the average of 5 weightings. In the posttest, however, only 76.54 % responds correctly. The McNemar change test shows a significant decrease, S(1, N = 162) = 10.0000, p = .0022.

Graphical items

An analysis of the graphical item where students have to distinguish a sample and a sampling distribution shows that performance increases drastically from before to after the activity (from 65.97% at pretest to 79.33% at posttest). The McNemar change test shows the increase is significant, S(1, N = 143) = 15.1250, p < .0001. An examination of the explanations shows that students' reasoning about the difference between sample and sampling distributions is much more profound for posttest responses. In the posttest many more students indicate that an element in a sample is only one element and that an element in the sampling distribution is the mean of a sample with three elements.

As mentioned before, only posttest data are available for the second type of graphical items. For these two items combined, students in our study have an overall average score of 9.31 on 14 (s = 2.84). This result is comparable to the average of 9.6 (s = 2.66) and 7.6 (s = 2.84), observed by Lunsford, Rowell, and Goodman-Espy (2006) for students following a post-calculus probability course (n = 18) and a statistics course (n = 7).

To assess students' reasoning about the variability and shape of the sampling distribution as the sample size increases from small to large, responses to reasoning pairs are coded in different categories, such as *Correct* or *Good*. A *Correct* response for a particular item means that the student picked the correct sampling distributions for both sample sizes (small and large). A response is coded as a *Good* response when students – although they do not pick the correct sampling distributions – realize that sample size has a negative impact on sampling variability and that the sampling distribution become more normal as sample size increases. The percentage of *Correct* or *Good* responses to the two graphical items of the posttest is 53.7%. This is relatively high compared to the percentages delMas et al. (2004) reported for three different versions of the *Sampling Distribution Activity*. In their study, for one version of the SDA (n = 118) this percentage was 39.4%, for a second version (n = 154) 38.6% and for a third version (n = 94) 36.7%.

5. Discussion

This study reports empirical data on students' understanding of the concept of sampling distributions before and after the Sampling Distribution Activity (SDA) (Chance et al., 2004). Comparison of pre- and posttest scores show a substantive improvement of students' understanding of sampling distributions for a graphical item that is closely related to the exercises in the activity, but a status quo or even a deterioration for two other (contextual, non-graphical) items.

Students' visual understanding of sampling distributions and their characteristics, like variability and shape given a specific sample size, was measured with two graphical items only after the activity. Although only about half of the students showed good to perfect visual understanding, this result was high compared to the results of delMas et al. (2004). The reason for this difference can for instance be differences in student factors (e.g., field of study, mathematical background, previous experience with statistics), differences in context factors (e.g., pedagogical methods of the teacher, classroom organization), or differences in implementation of the SDA. Further study is needed to reveal the impact of these factors in more detail.

In general, this study confirms the complex and difficult nature of the topic of sampling distributions for students. Because of a better performance for graphical items compared to non-graphical items, it seems to indicate that the potential of the SDA is highest for visual understanding of sampling distributions and that transfer from the simulation software to other representations or contexts is most difficult. We agree with the designers of the activity that software alone cannot build reasoning about sampling distributions. The study suggests that students might benefit from even more diversity in the activity (for instance, different visualizations or contexts) than currently available.

REFERENCES

Chance, B., delMas, R. C., & Garfield, J. (2004). Reasoning About Sampling Distributions. In D.Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). The Netherlands: Kluwer Academic Publishers.

delMas, R. C., Garfield, J., & Chance, B. (2004). Using assessment to study the development of students' reasoning about sampling distributions. *Paper presented at the annual meeting of the American Educational Research Association.* San Diego, CA.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2,* 22-38.

Lunsford, M. L., Rowell, G. H., & Goodman-Espy, T. (2006). Classroom Research: Assessment of Student Understanding of Sampling Distributions of Means and the Central Limit Theorem in Post-Calculus Probability and Statistics Classes. *Journal of Statistics Education, 14 (3),* Retrieved April 2, 2006 from http://www.amstat.org/publications/jse/v14n3/lunsford.html.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education, 10(1),* Retrieved February 10, 2006 from http://www.amstat.org/publications/jse/v10n1/mills.html.

Moore, D. S., & McCabe, G. P. (2006). *Introduction to the Practice of Statistics: fifth edition.* New York: W.H. Freeman and Company.

Tversky, A., & Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. *Science, 185,* 1124-1131.