12th International Congress on Mathematical Education
Program Name XX-YY-zz (pp. abcde-fghij)
8 July – 15 July, 2012, COEX, Seoul, Korea

SUPPORTING STUDENTS TO DEVELOP CONCEPTS UNDERLYING SAMPLING AND TO SHUTTLE BETWEEN CONTEXTUAL AND STATISTICAL SPHERES

Adri Dierdorp, Arthur Bakker, Jan van Maanen, & Harrie Eijkelhof

Freudenthal Institute for Science and Mathematics Education

a.bakker4@uu.nl

To stimulate students' shuttling between contextual and statistical spheres, we based tasks on professional practices. This article focuses on two tasks to support reasoning about sampling by students aged 16-17. The purpose of the tasks was to find out which smaller sample size would have been sufficient for making reliable inferences. The research question addressed is: How can students be supported to develop concepts underlying sampling and to shuttle between contextual and statistical spheres? Design research was carried out to test whether the tasks had the potential to support students' concepts underlying sampling and to find indications of what teachers should do to use this potential. Analysis of video recordings indicates that the students showed a balanced development of the concepts underlying sampling. They seemed aware of the purposes of the tasks and were able to apply their statistical knowledge, but tended to forget to shuttle back.

Keywords: sampling, inferential reasoning, authentic practice, purpose, utility.

Shuttling between contextual and statistical spheres

This paper deals with a fundamental challenge in educational design: to stimulate students to shuttle back and forth between a contextually phrased problem and statistics. The importance of moving between the contextual world of life and the statistical world of symbols is stressed in many different areas. Many contextual problems can be solved with the help of statistics. However, the results of computation or modeling need to be evaluated on their merits and validity in the context. More generally, Wild and Pfannkuch (1999, p. 228) stressed the significance of "shuttling between the contextual and statistical spheres."

From the literature, we know that such shuttling is not easily promoted in students. Ainley, Pratt, and Hansen (2006) propose to focus on the purpose of tasks and utility of what is learned. A purposeful task is one that has a meaningful outcome for the student, for example in terms of an engaging problem. Utility is the construction of meaning for the ways in which mathematical concepts are useful. Ainley et al. (2006, p. 25) consider "the provision of authentic tasks inherently problematic," but we think that suitable authentic contexts can provide both purpose and utility. The approach we investigate in this paper is to base tasks in upper secondary education on problems from authentic professional practices. Research has shown that such a design approach can help students see the purpose of what they do in classrooms and the utility of what they learn. However, such an approach might come at the expense of conceptual learning.

The aim of this paper is to examine whether realistic tasks, inspired by authentic professional practices, and embedded in appropriate teaching, can in principle support students'

conceptual understanding of sampling in such a way that they also learn to shuttle between contextual and statistical spheres. To us, this "shuttling" not only includes students' application of statistical knowledge, but also seeing the purpose of the task, and the utility of what they learn.
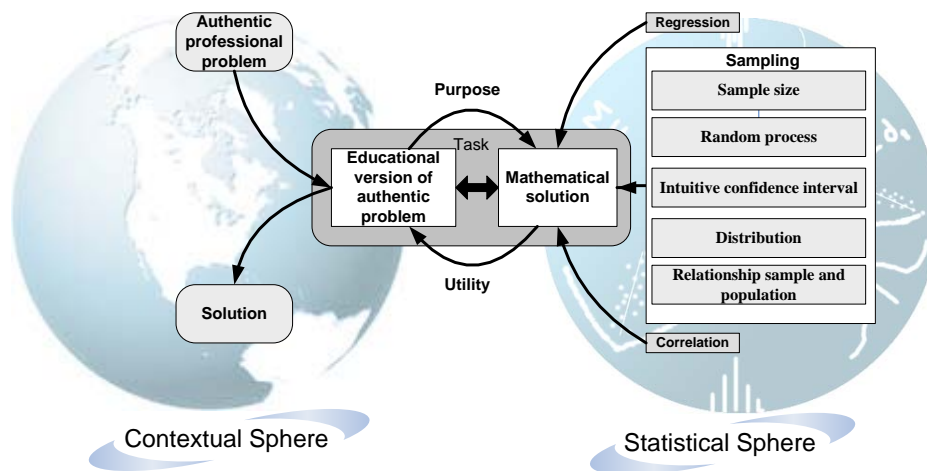
CONCEPTS UNDERLYING SAMPLING

Figure 1. Shuttling between the contextual and statistical spheres during the sampling tasks.

Basing tasks on situations from authentic professional practices has been studied in science education, but much less so in statistics education (Dierdorp, Bakker, Eijkelhof, & Van Maanen, 2011). Here we focus on sampling to help students see the utility of correlation and regression and shuttle between the contextual and statistical spheres. Sampling is considered a key aspect to the teaching of informal inferential reasoning. For example, Pfannkuch (2008, p. 1) argued: "when students are not aware of sampling their informal inferential reasoning is limited." We address five concepts underlying sampling (inspired by Pfannkuch, 2008, p. 4) which are important for students' statistical reasoning: sample size, random process, distribution, intuitive confidence interval, and relationship between sample and population. We address these five concepts one by one and show later that we recognize these concepts in students' reasoning. The paper's focus is schematically represented in Figure 1.

Sample size and law of large numbers

Students need to understand that increasing sample size generally leads to better estimates of probability and population characteristics. For understanding sampling it is therefore necessary that students develop a concept of sample size. A "big idea" connected with sample size is the law of large numbers which says that predictions become more reliable when made from larger than from smaller samples.

Random process

Students also need to realize that random processes such as repeated measurements of the same phenomenon will lead to different outcomes. Then they can understand that inferences are influenced by the drawn sample.

Distribution

The aforementioned concepts underlying sampling relate to the big idea of distribution. Rubin, Bruce, and Tenney (1990) argued that a sample gives the practitioner information about the distribution of a population and that this is a central idea of statistical inference. Drawing enough samples can support students' understanding that the shape of the graph obtained by a "bigger" sample becomes more similar to the graph of the population as a whole. Confronting students with samples of increasing sizes, so called "growing samples," can help students to become more aware of emerging distributions by means of stabilizing measures of variation (tendency), and smoothening shape (e.g., Bakker, 2004).

Intuitive confidence interval

Shaughnessy (2006, p. 87) argues that students "should have a sense of the reasonably expected variability around the expected value, something as a confidence interval". In real life, predictions are not based on one value obtained by a regression line because often a margin around the predicted value is essential. For example when a physiotherapist finds a client's peak heart rate just under the value predicted by a common formula he or she will not worry (Dierdorp et al., 2011). He has a sense of what could be called an intuitive confidence interval.

Relationship of sample and population

Research into school statistics reports the problems students have in drawing inferences that make sense in the context (Makar & Rubin, 2009). One problem of how to draw sensible inferences is caused by a lack of awareness of variability when generating samples. Samples often provide a distorted image that is yet in some way representative.

Research question

Given the problems and challenges mentioned in Section 1, and the complex multi-faceted concepts underlying sampling summarized in Section 2, we formulated the following research question: How can students be supported to develop concepts underlying sampling and shuttle between contextual and statistical spheres?

METHOD

The research question is addressed through two case studies that were part of a design experiment on correlation and regression with a twelfth-grade group of thirteen students and an eleventh-grade group of sixteen students from the pre-university track. Both groups had opted to study the school subject "Nature, Life, and Technology". The first group was taught by the first author. The second group was taught by another teacher, with the first author observing and interviewing. Two sample tasks covered three of the 23 lessons, each 50 minutes in both schools. Students used Fathom for drawing samples and Excel for making their own scatter plots to investigate their results from the sampling software.

Sampling   tasks
To stimulate students to develop concepts underlying sampling and to shuttle between contextual and statistical spheres we designed two realistic tasks based on authentic professional practices.

To stimulate students' reasoning about sampling we drew on a professional practice, namely research on peak heart rates (PHR). Gellish et al. (2007) measured many people and found a different relationship (PHR = 207 – 0.7A, with A as age) between age and peak heart rate than the one typically used in sport physiotherapy (PHR = 220 – A). We provided the students with their data set of 908 measurements. In the Heart rate task we asked the students if we could do with a smaller sample: What smaller sample size would be sufficient to find a reliable formula that is close to the original formula based on Gellish et al.'s data set?

The task is based on the instructional idea of growing samples. Such an approach is helpful in supporting coherent reasoning about sampling, distribution and other statistical key concepts (Bakker, 2004). The students started with Gellish et al.'s data set and regression line, and then investigated smaller samples to find the smallest sample size that still produced a reliable model using the sampling option in Fathom. We conjectured that the question about how small a sample can be and still allow a sufficiently reliable inference can stimulate students to reason about all five aforementioned sampling concepts (and not just one aspect of sampling at the time, as happens in atomistic approaches to task design).

The second task is inspired by the practice of monitoring the height of dikes. A dike is an artificial construction to prevent flooding. Dike monitoring is essential for the Netherlands because large parts of the country are below sea level. A persistent problem is that dike heights decrease over time. If the height reaches a "critical value" (so named by the Ministry of Transport and Water Management) high sea and river water levels are a danger.

We introduced the Dike sampling task: Keeping in mind the high cost of measurement, what smaller sample would still have led to a reliable prediction of when the critical value will be reached? The students got 44 real data points of the deviations of a dike location.

The students could change the sample size and got the corresponding scatter plot with a regression line and a formula. They had to decide themselves which number of measurements was required to find a reliable prediction based on a smaller sample which is close to the prediction obtained by the regression line for the complete set. The students were only told that they had to save money by reducing the number of measurements.

To address the research question we sought an efficient way to check if our tasks could in principle support students' concepts underlying sampling and their shuttling between context and statistics. To this end it seemed sufficient to use case studies of students working with these tasks.

The first case study focused on Rick, aged 17. While Rick worked on the tasks, we video recorded his activities and transcribed the spoken text. To identify which sampling concepts were at stake in the interaction, we divided the transcripts into 26 fragments of interactions between students or between student(s) and teacher. Each fragment consisted of several turns, i.e. the spoken text of a person which is not interrupted by another person as a turn. Each fragment was coded by means of interpretive microanalysis with one or more items of the five aforementioned concepts underlying sampling.

The second case study focused on two 11th grade students, Sean and Kars, both aged 17 from another yet similar school. Sean and Kars were both motivated students, and their school

results were similar to Rick's. Sean and Kars wrote in their questionnaires that they were not familiar with the authentic professional practices or the statistical techniques central in the tasks. This second case study focuses on the Dike sampling task, during which the researcher observed and interviewed the students. We video recorded the interaction and transcribed the spoken text. To analyze what teachers may need to do to stimulate students to shuttle between the contextual and statistical spheres we divided the transcripts into three phases, and categorized the researcher's attempts to help them shuttle back from statistics to context.

RESULTS

We present the results of the analysis of the data collected during the Heart rate and Dike sampling tasks and show what kind of reasoning about sampling these tasks potentially support.

Developing concepts underlying sampling (Case 1)

Case study 1 was carried out to investigate the potential of the tasks to support students' development of concepts underlying sampling. We first present a quantitative impression.

Table 1: Concepts underlying sampling recognized in each fragment of Rick's spoken text.

	Fragment	Sample Size	Random Process	Distribution	Confidence Interval	Relation Sample/Population	Time (seconds)	# Turns
Heart rate task	1	x					50	6
	2	x	x				36	5
	3			x			41	4
	4	x					24	3
	5	x		x			48	5
	6					x	56	9
	7	x	x				80	8
	8			x		x	56	7
	9	x		x			48	5
	10				x		71	9
	11		x				59	15
	12					x	13	2
	13	x	x	x			46	6
	14	x			x		67	6
	15	x			x		65	6
Dike task	16					x	20	4
	17		x	x		x	55	7
	18				x	x	15	3
	19			x		x	35	4
	20				x	x	20	4
	21					x	15	4
	22		x		x		15	2
	23		x	x			25	5
	24	x				x	40	7
	25	x		x	x	x	40	7
	26					x	50	6
	Total	11	7	8	7	12		

From the rather well balanced distribution of codes across the different concepts underlying sampling, we conclude that the tasks have the potential to raise students' concepts underlying samples. Many fragments were coded by more than one concept suggesting that these concepts were developed and used in relation to each other. In the following subsections we illustrate the five conceptual concepts underlying sampling. The origin of excerpts is indicated by Fn, where n is the number of the fragment.

Sample size

From the eleven fragments in sample size column of Table 1, we infer Rick had developed a sense of sample size and law of large numbers. For example, he expected that when the sample size becomes big enough the regression coefficients would stabilize. He also formulated it reversely as follows:

> Rick [F4]:   You get [when drawing 100 instead of 500 cases] more variability. More deviations but the line will have a negative slope. The intersection with the vertical axis will be different too. By changing the number of cases you can monitor the effect on the regression line.

This last sentence may indicate that he understood the utility of sampling. In other words: he seemed to understand why the concept of sampling is useful in this task. When he drew samples of size 50 he mentioned the difference of the coefficients of the regression lines when executing samples of this same size. He mentioned that when he drew bigger samples, he would expect a formula closer to the original formula of the population. Such examples indicate that Rick was aware of the effect of sample size and had a sense of the law of large numbers.

Random process

There were indications that Rick developed understanding of the concept of random process during his work on the task. At first, he believed that a bigger sample size would lead to a more reliable regression line, but when he took a sample of size 100 he noticed that the regression coefficients deviated more than those of the last sample of size 50. He then realized that a relatively small extension of sample size would not necessarily lead to a "better" formula and tried to explain this with the concept that the software randomly sampled the higher peak heart rates. The teacher tried to stimulate a reflection on Rick's former statement that a higher amount of cases would imply a formula for the regression line more similar to the original. Rick seemed to be clear about the fact that the randomness of the sample leads to unpredictable outcomes and that it is not necessary [F7] to find a regression line more similar to that of the larger set when taking a slightly bigger sample. We see this as growth in Rick's understanding of how samples behave.

Intuitive confidence interval

Rick mentioned that with bigger samples the points were getting closer together. He entered all regression coefficients in a spreadsheet, seemed to be aware of an interval per sample size, called this "margin," and found a margin found for samples size of 200 and up acceptable to make a prediction. Rick considered a margin around an expected value: "Till 200 it is too varied, if you do not combine. It can be a coincidence. You can build a safety margin. Then you go to 250." He explained that the margin of the slopes he found at size 200 could be small as the result of coincidence and thought that a sample size of 200 would allow him to predict a reliable regression line. When he said "if you do not combine" he probably meant that this is only the case when focusing only on the slope. For safety reasons he suggested a sample size which was one "step" bigger, such as 250. The fact that Rick mentioned a safety margin

supports that the realistic context of this task may have provoked him to reason about the outcomes.

Distribution

In previous lessons the students discussed the role of a physiotherapist using a regression line to advice his clients. Rick and his fellow students decided that a client's peak heart rate must be in a margin around the regression line. When the teacher asked Rick if he expected a larger correlation when drawing a larger sample, he answered negatively and explained:

> Rick [F13]: When you start taking random values and do so the tenth time, taking random values, they are still random. Only the margin will become more colored [filled by colored dots]. It will not be wider and will never become narrower and never become much wider.

This excerpt indicates Rick's sense of margin and distribution and that he seemed to be aware of the purpose of the task. He expected the margin to be about the same for each sample size. Only when drawing a sample with a bigger size did he mention that the margin became more colored. Probably he mentioned this because he expected more points within the margin when the sample is larger.

Rick considered the shape of the distribution (see Figure 2) which he called "a trumpet shape." From his remarks it was clear that he had expected to see a stabilizing trumpet shape. He saw this shape in the Dike sampling task as well. This time he did not draw the trumpet, but only mentioned: "You see again the trumpet shape". At this point we proceed with the Dike sampling task, in which students were challenged to apply what they had learned in the Heart rate task.
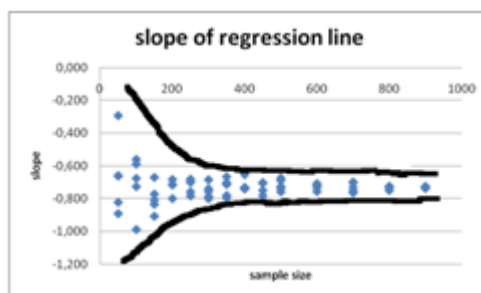

Figure 2. The trumpet shape distribution of the slopes of regression lines obtained by growing samples drawn by Rick.

Relationship between sample and population

Rick showed some understanding of a relationship between sample and population and seemed to be aware of the purpose of the task: to find a smaller sample size in order to save money and still have a reliable formula for the regression line. He said to his fellow student Eline: "It [software] plots the regression line. Then you are able to see how much points [measurements] can be saved and still find a reliable regression line [F18]." Further analysis suggests that Rick was aware of the purpose of the task and the utility of sampling. Rick talking about the reliability of a sample with a certain size indicates that he thought about the relationship between the sample and population distribution. He compared the results of

every sample with the results of the original set. The observations gave us indications that other students also developed conceptual understanding of the relation between the sample distribution and population distribution, but shuttling back to the context was not obvious to them. Many only considered the trumpet shape distribution of the slope to consider the sample size for making a reliable inference. Only Rick and Eline shuttled between the spheres basing their inference on the total effect on the deformations instead of only looking at the slope of the regression line.

Using the Dike sampling task's potential to improve students' shuttling (Case 2)

To study how a teacher can use the task potential to improve students' shuttling between the statistical and contextual spheres, we carried out another case study. From this case study we present the Dike sampling task, in which the researcher needed to put a great deal of effort in supporting Sean to link back to the context. We distinguished three phases in Sean's working on the task with his fellow student Kars: Phase 1: in which students were self-reliant working on the task, Phase2: with several types of questions and hints by the researcher, and Phase 3: in which the students were self-reliant working on the researcher's context question.

During the first twenty minutes (Phase 1) Sean and Kars worked self-reliantly. Although all tasks were based on authentic professional contexts and the students seemed to see the purpose of the task, they stayed in the math world focusing on formulas:

Sean: The slope [-0.00124, slope of the regression line at sample size 20] is almost the same [as from the original formula, -.00123].

Kars: Yes, the b too [the b from y = ax+b; -2.74 vs. -2.8].

After about twenty minutes, when the researcher discovered that Sean and Kars reasoned without referencing back to the contextual problem, he tried to focus their attention to the contextual meaning of their decisions in several ways (Phase 2). To gain insight into the support that the researcher gave, we classified his turns as informative questions, with which he tried to find out what the students did and meant (n = 10), revoicing questions and remarks in which he rephrased the students remarks (n = 7), explanation questions to find out why they did certain actions (n = 5), instructional support to do general suggestions (n = 2), and reducing degrees of freedom support to let the students shuttle between the contextual and the statistical sphere (n = 1). Despite the researcher's questions, Sean and Kars stayed focused on the slope of the regression line and did not use the context. Often the slope was different only in the fourth decimal, and they thought that the formula obtained by sample size 30 was close enough to the original formula. They judged this purely on the basis of the formula, whereas the researcher hoped they would think through the contextual consequences of the differences between the regression formulas.

At the end of phase 2, the researcher again tried to reduce the degrees of freedom" and asked more specifically about the difference in days between the prediction based on the regression line with sample size 30 and the prediction based on the original formula based on a sample of 44 (Phase 3). Sean and Kars then continued to work self-reliantly on this question. They used the spreadsheet to calculate the day for which according to the regression line H = -10, and found that the random sample with size 30 produced a regression line which would imply

raising the dike almost a year later than the original line. Then they realized that the context asked for a more precise approach. Because Sean and Kars were no longer satisfied with sample size 30, they decided to take samples of other sizes. For each sample they also calculated the difference between the corresponding calculated predicted days and the day when the original formula would predict the critical value of -10 mm. They also calculated the average of differences for each sample size. They decided that in this context sample size 40 was acceptable. They were disappointed that they only saved four measurements, but when the researcher asked how much money would be involved in skipping four helicopter flights, they were more satisfied. This last case study indicates that the students were focused on the mathematical concepts of the tasks. The researcher repeatedly had to emphasize the contextual problem to stimulate the shuttling back to the contextual sphere. He had to reduce the "degrees of freedom" by asking about specific contextual consequences of difference in the formula obtained by their sample and the original formula.

CONCLUSION AND DISCUSSION

Our research question was how students can be supported to develop concepts underlying sampling and to shuttle between contextual and statistical spheres. It seems possible to use authentic problems from professional practices to design tasks that are purposeful from a student perspective and lead students to see the utility of what they learn. This might help students apply what they have learned. However, it is not self-evident that students develop rich conceptual understanding from authentic tasks because designers seem to have less control about what students learn conceptually. We wanted to know whether it is feasible, in principle, that students see purpose and utility while also developing a rich conceptual understanding of what is at stake, in this case sampling. In response we conclude that the analyses show that the realistic sampling tasks, inspired by authentic professional practices, are rich and focused enough to stimulate reasoning about the concepts underlying sampling in a balanced way and in relation to each other. This seems an advantage over atomistic approaches to statistics education that deal with aspects of concepts one by one (cf. Bakker & Derry, 2011) and this seems to address Ainley et al.'s (2006) concern that engaging tasks seem to be less focused.

Because with task design the devil is in the detail, it is worth speculating on what makes the sampling tasks suitable in helping students develop the main concepts underlying sampling in relation to each other. We think this has to do with the question of asking how small a sample can be so that the inference is still reliable. This explicitly requires reasoning about sample size and the relation between sample and population. Moreover, the issue of randomness comes up when students compare samples with the same size when trying to judge if inferences are reliable. When comparing samples of the same size, intuitive confidence intervals and distributions also come into play. These findings are in line with earlier findings at the middle school level that tasks based on growing samples have the potential to stimulate students to reason about multiple facets of distribution and uncertainty.

We also wanted students to shuttle between contextual and statistical spheres. More concretely, we think it is important they learn to model contextual problems and apply their statistical knowledge, but also focus on purpose and utility. The purpose of the tasks was clear

for the students involved: to find a smaller sample size in order to save money and still have a reliable formula for the regression line. It seemed that they did see the utility of sampling in order to find such a reliable formula.

By means of a second case study we explored what types of questions teachers may need to ask. What helped best was to ask specific questions about practical consequences of mathematical issues (e.g., what would such a small difference in coefficients mean in terms of the prediction?).

We addressed a persistent design challenge and do not claim to have solved it. More research is needed to investigate the support students need to shuttle between contextual and statistical spheres. More specifically, we suggest investigating teachers' scaffolding of students' shuttling.

References

Ainley, J., Pratt, D., & Hansen, A. (2006). Connecting engagement and focus in pedagogic task design. *British Educational Research Journal, 32*(1), 23-38.

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal, 3*(2), 64-83.

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning, 13,* 5-26.

Dierdorp. A., Bakker. A., Eijkelhof.**,** H. M. C., & van Maanen, J. A. (2011). Authentic practices as contexts for learning to draw inferences beyond correlated data. *Mathematical Thinking and Learning, 13,* 132-151.

Gellish, R. L., Goslin, B. R., Olson, R. E., McDonald, A. R. G., Moud, G., & Virinder, K. (2007). Longitudinal modeling of the relationship between age and maximal heart rate. *The American College of Sports Medicine, 39,* 822–829.

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

Pfannkuch, M. (2008). Building sampling concepts for statistical inference: A case study. In *Proceedings of the International Congress on Mathematical Education*, Monterrey, Mexico, 6-13 July, 2008.

Shaughnessy, J. M. (2006). Reseach on students' understanding of some big concepts in statistics. In G. F. Burrill & P. C. Elliott (Eds.), *Thinking and reasoning with data and chance* (pp. 87-98). Reston, VA: The National Council of Teachers of Mathematics, Inc.

Westbroek, H., Klaassen, K., Bulte, A., & Pilot, A. (2010). Providing students with a sense of purpose by adapting a professional practice. *International Journal of Science Education. 32,* 603-627.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *Statistical Review, 67,* 223–265.