

ELEMENTARY PRESERVICE TEACHERS' REASONING ABOUT MODELING A "FAMILY FACTORY" WITH TINKERPLOTS – A PILOT STUDY

ROLF BIEHLER

University of Paderborn
biehler@math.upb.de

DANIEL FRISCHEMEIER

University of Paderborn
dafr@math.upb.de

SUSANNE PODWORNÝ

University of Paderborn
podworny@math.upb.de

ABSTRACT

Connecting data and chance is fundamental in statistics curricula. The use of software like TinkerPlots can bridge both worlds because the TinkerPlots Sampler supports learners in expressive modeling. We conducted a study with elementary preservice teachers with a basic university education in statistics. They were asked to set up and evaluate their own models with TinkerPlots by using a real and open dataset they were given. In this article we present students' processes of setting up and evaluating their models and focus on their reasoning during this process.

Keywords: *Statistics education research, Preservice teacher education, Data modeling*

1. INTRODUCTION

Data and chance have received increased attention in Grades 3 and 4 in mathematics classrooms in Germany (pupils aged 8-10). This requires a greater focus on data and chance in mathematics teacher education as well. At the University of Paderborn we designed a course "Modeling, Magnitudes, Data and Chance" (MMDC) for mathematics preservice teachers at the primary school level to expand their statistical content knowledge about exploring data and simulating chance experiments. We used TinkerPlots (Konold & Miller, 2011) as a digital tool. A primary focus in our curriculum is to enter the worlds of data and chance and relate them to one other. We do this via simulation of chance experiments. In this paper, we report on a laboratory study with a small number of randomly chosen students from the MMDC course. These students were asked to work on a task bringing together data and chance. We conducted clinical interviews some weeks after the course to investigate the cognitive processes of the participants while building, simulating, and evaluating a model with TinkerPlots for real data on the gender distribution in families.

2. RELEVANT LITERATURE

We briefly review literature on three relevant topics: The metaphor of a “data factory” for connecting data and chance, the notion of expressive modeling, and background literature on the data set we used in the study.

Konold, Harradine, and Kazak (2007) introduced the idea of using the TinkerPlots Sampler in a specific way to bring together data and chance. They used the metaphor of the Sampler as a “data factory” that produces “real objects” carrying several attributes. A simulation is thus used not only for modeling a single attribute and its distribution, but also for modeling and reproducing bivariate or multivariate data.

This approach goes beyond traditional contexts for probability in school and makes students build and refine data-producing factories. Konold et al. (2007) studied this approach with twelve students, aged 12-14. The students took part in a yearlong after-school program exploring probability. Benefits of this approach were reported. Students quickly engaged with the task and the activities helped them to develop beginning ideas of statistical inference. This goes together with informal inferential reasoning (Makar & Rubin, 2009; Pratt & Ainley, 2008), because it involves making generalizations from samples without the use of formal statistical tests. The “first model is a guess, an expectation, or prediction” (Konold & Kazak, 2008, p. 2) and is refined later in the process when comparing the simulated data to the real data. The students looked at an empirical distribution and built a model with the Sampler to reproduce the distribution. The computer was used to see how well the simulated results fit the empirical distribution. An in-class discussion was facilitated about the characteristics of a good model. The discussion and its results are not reported in detail by the authors. Konold and Kazak introduced four main ideas connecting data and chance that emerged from their analyses of students’ conceptions: model fit, distribution, signal and noise, and the law of large numbers. They put emphasis on the aspect of controlling noise by adjusting the sample size.

According to Garfield and Ben-Zvi (2008), two main uses of statistical models can be distinguished:

Select or design and use appropriate models to simulate data to answer a research question. [...] Fit a statistical model to existing data or data that you have collected through survey or experiment in order to explain and describe the variability. (p. 144)

For our approach to modeling, the notion of “expressive modeling” is helpful. Pratt, Davies, and Connor (2011) emphasize that “in EDA, students express their own informal models for the data by searching for trends and patterns in the data, a process often referred to as expressive modeling” (p. 99; see also Doerr & Pratt, 2008). These authors distinguish between using models and building models and further distinguish between exploratory modeling and expressive modeling, citing Bliss and Ogborn (1989): “This distinction between using models and building models has also been described as the difference between exploratory modeling and expressive modeling” (Bliss & Ogborn, as cited in Doerr & Pratt, 2008, p. 264). Doerr and Pratt (2008) emphasize that “exploratory models are those models that are constructed by experts to represent knowledge in some content domain. Learners typically explore consequences of their actions within the boundaries of these content domain models” (p. 7). In contrast to exploratory models, they characterize expressive models in the following way:

Building a model (or expressive modeling) provides learners with the opportunity to express their own concepts and to learn through the iterative process of representing their ideas, selecting objects, defining relationships among objects, operating on those relationships, and interpreting and validating outcomes. (Doerr & Pratt, 2008, p. 8)

Doerr and Pratt (2008) also mentioned that expressive modeling:

[...] is a significant shift in perspective from the activity of exploring a pre-built model that necessarily embodies the intentions, concepts, and structures of an expert. The process of model building forces students to make explicit their own ideas about the relationships among objects and variables to examine, interpret, and validate the consequences of their ideas. (p. 8)

In addition and consequently to the statements above, TinkerPlots can be a valuable software to support this process. Pratt et al. (2011) pointed out that

new developments in TinkerPlots promise to provide a graphical probabilistic language to model the generation of data sets (Konold et al., 2007). Teachers could use the software as an authoring tool in which they build models for students to explore or as an expressive tool in which students build their own models of phenomena. (p. 99)

In our case TinkerPlots was used as an expressive tool in the sense of the definition of Pratt et al.

In Germany there is a long-standing tradition of connecting data and chance in the curriculum that is named *stochastics* instead of *probability and statistics* (see Burrill & Biehler, 2011). Several examples have been suggested for classroom use but not explored by an accompanying empirical study. One such example is the so-called Geissler data. Data on gender distribution of children in families with up to 12 children in Germany in 1880 were collected by A. Geissler (1889), who was a German doctor and statistician in the nineteenth century. The data set became well-known internationally. Fisher, one of the major statisticians at the beginning of the 20th century, explored Geissler's data for families with eight children and found some irregularities compared to expected values of a model based on the binomial distribution. He explained some of the irregularities with the occurrence of multiple births but left open unexplained deviations to future studies (Fisher, 1970). Biehler (2005) proposed using Geissler data in the classroom by concentrating on one aspect of the data: the number of boys in 10690 families with 12 children each. He used the binomial distribution as a theoretical background and showed some interesting aspects of the data and the modeling process using different models, informal statistical reasoning, and statistical tests. Taking the binomial distribution with $p = 0.5$ for a boy's birth as a model, "the residual display shows a systematic deviation that can hardly be interpreted as random fluctuation" (p. 8). Calculating the overall percentage of boys in all 10690 families as $p = 0.5168$ and using this as a model for a boy's birth leads to smaller deviations, but they are still systematic deviations. Biehler pointed this out as an important finding: "Having discovered systematic deviations from the simple binomial model is a genuine discovery" (p. 9). The binomial model alone may not explain the data. The data were also used in an American textbook. The authors stated, "The systematic pattern of deviations from the binomial distribution suggests that the observed variation among families cannot be entirely explained by the independent-trials model" (Samuels & Witmer, 2003, p. 113). Further explanations are necessary and may be discussed with students. Other uses include in several German textbooks (Griesel, Postel, Suhr, & Gundlach, 2003; Harten & Steinbring, 1984), but an empirical study on students' reasoning has never been done. We aimed to use the modeling approach with these interesting data to gain insights about preservice teachers' informal inferential reasoning with models.

3. BACKGROUND FOR THE STUDY: THE MMDC COURSE FOR PRESERVICE TEACHER EDUCATION

The MMDC course consists of two parts: data analysis and introduction to probability. In the data analysis part, the preservice teachers experience an entire PPDAC-cycle (Wild

& Pfannkuch, 1999). At the beginning a statistical question is generated out of a statistical problem (first “P” in the PPDAC-cycle) followed by planning of data collection (e.g., setting up a questionnaire for collecting data, the second “P” in the PPDAC-cycle). After these data are collected (“D” in the PPDAC-cycle) the data are analyzed (“A” in the PPDAC-cycle) with TinkerPlots. Finally conclusions (“C” in the PPDAC-cycle) and interpretations of the findings of the data exploration process are documented (e.g., in the form of a report). Comparing groups is a fundamental aspect. In the second part of the course, the participants are introduced to classical probability and the frequency interpretation of probability via the law of large numbers, including the so-called $1/\sqrt{n}$ law (see below). Chance experiments are modelled with TinkerPlots throughout the course. Along with simulating chance experiments, preservice teachers learn about the accuracy of simulations in the sense of a “rule of thumb” (distance between the probability and the relative frequency) for several numbers of repetitions ($n = 100, 1000, 5000, 10000$). Finally, the preservice teachers learn basics about the binomial distribution and get first insights into informal hypothesis testing. Table 1 summarizes the content of the course.

Table 1. Content of the MMDC course

No. of lecture	Content of MMDC course
1	Problem and Plan: Collecting data; Basic terms of descriptive statistics
2	Displaying data with TinkerPlots: pie charts, bar charts, value bars, describing and interpreting displays; calculating median and mean of distributions of numerical variables; displaying median and mean in TinkerPlots
3	Absolute and relative frequencies, calculating quartiles of numerical variables, displaying histograms and boxplots in TinkerPlots; describing and interpreting histograms and boxplots
4	Describing and interpreting range and interquartile range, skewness and shift; group comparisons I (two categorical variables)
5	Group comparisons II (one categorical and one numerical variable)
6	Group comparisons II (one categorical and one numerical variable) & group comparisons III (two numerical variables)
7	Combinatorics
8	Probability: Basic terms and definitions, historical glimpse, simulating chance experiments
9	Empirical law of large numbers, $1/\sqrt{n}$ law, accuracy of simulations; “rule of thumb”
10	Galton board, binomial coefficient and Pascal’s triangle, binomial distribution I
11	Binomial distribution II, Bernoulli process
12	Informal hypothesis testing mp3 vs. CD quality test (cf. Riemer, 2009) in analogy to the “lady tasting tea” (see Fisher, 1971)

The preservice teachers experienced a broad spectrum of data and chance contexts. The connection of data and chance was predominantly done by connecting relative frequencies and probabilities. On the one hand the students were taught how to estimate probabilities by relative frequencies of events in chance experiments and on the other hand they were taught how to predict relative frequencies with probabilities of events in chance experiments. We also taught them to judge the accuracy of simulation (accuracy of estimating the probability of an event of a chance experiment) by using the $1/\sqrt{n}$ law. We

handed out a table (as a “rule of thumb”) with several sample sizes and the corresponding 95% accuracy interval of the simulation on the background of the $1/\sqrt{n}$ law (Table 2).

Table 2. Accuracy of simulation as “rule of thumb”

Sample size	95%-accuracy (radius of the interval)
50	0.140
100	0.100
1000	0.030
5000	0.015
10000	0.010

Both directions of the $1/\sqrt{n}$ law were taught to the preservice teachers. At first they learned that the $p \pm 1/\sqrt{n}$ law is predicting the interval where the relative frequency of an event can be expected with 95% probability, if the probability p is known. The second direction of the $1/\sqrt{n}$ law is to make inferences about the unknown probability p , when a relative frequency of h_n of an event is observed. We communicated the rule that we estimate p in the interval $h_n \pm 1/\sqrt{n}$ and specify the certainty of this estimation as 95%. We stopped at this level of informal confidence intervals and did not go into more details. Moreover, we did not teach our students explicitly about fitting models to given (empirical) data nor formal tests of goodness of fit.

4. THE STUDY

4.1. RESEARCH QUESTIONS

Setting up a data-based model, evaluating it with regard to real data, and potentially adapting the model in a second step, was a new process for our participants. We were interested in the following research questions:

1. To what extent are our preservice teachers able to solve the Geissler task successfully?
2. Which phases and overall structure can be identified when preservice teachers are working on the Geissler task with TinkerPlots?
3. In which circumstances do the preservice teachers need support?
4. How do the preservice teachers reason in detail about their models in the face of real data?

4.2. PARTICIPANTS

We conducted the interview study with 14 randomly selected participants from the MMDC course about eight weeks after the course ended. The participants were asked to solve a modeling task with TinkerPlots in pairs. We had seven pairs in all. The participants were in their second year of teacher education at university level to become elementary school teachers.

4.3. THE GEISSLER TASK

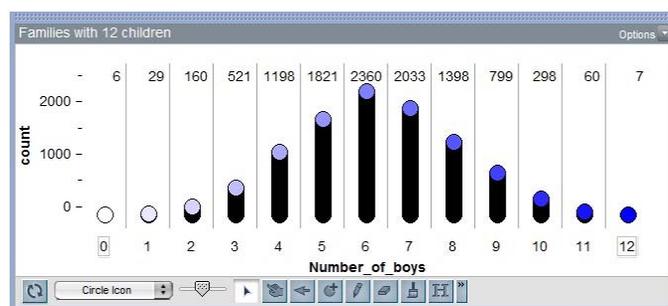
A major point in the design of the task was to use real data, where plausible, to apply the binomial model to what the students knew from the course. We used the “Geissler data” (Geissler, 1889), a dataset that contains, among other things, the number of boys in 10690

families of 12 children each living in Saxony in 1880. This task of comparing whole distributions to real data was new for our preservice teachers. Therefore, we designed several interventions (for details see interview procedure (subsection 4.5) and Appendix C) on the basis of our partial task analysis (subsection 4.4) to support the participants when working on the task. The exact wording of the task given to the participants can be found in Figure 1. The students also received a TinkerPlots file with the data (Figure 2).

Large families in the 19th century

TinkerPlots as family factory?

From 1876 until 1880, A. Geissler collected demographic data about families in Saxony. He took into account five million births in Saxony. In this task we will focus on families who have twelve children each. The dataset contains data of 10690 families with a total of 128280 children. The following figure shows the distribution of number of boys in 10690 families.



Main question: *Does the TinkerPlots Sampler offer the possibility to model the distribution of gender in 10690 families with a random process, so that the distribution produced by this random process is nearly identical to the distribution of boys in the empirical data?*

Part 1: Understanding the distribution

What is the meaning of “1821” in the bin “5 boys”? Describe the shape of the distribution. Are you surprised about the distribution?

In the MMDC course you were introduced to several applications of TinkerPlots. One of these applications was to consider the TinkerPlots Sampler as a “factory”, which produces specific data. In the following sub-tasks we want you to set up a model that works as a family factory and produces families with 12 children.

Part 2: Setting up a family factory in TinkerPlots

Try to set up a model in TinkerPlots that reproduces the distribution of gender of the families with twelve children in Saxony. Is it possible to produce a distribution via simulation that is nearly identical to the distribution above?

Part 3: Comparing distributions

Compare the simulated and empirical distributions. What do you see? Does the simulated data fit the Geissler distribution?

Part 4: If you are not satisfied with your comparison, adjust your model and simulate again.

How can you change your assumptions of part 2 so that the data produced by the model fit better to the empirical data?

Finally we ask you to answer the question from the beginning: Does TinkerPlots Sampler allow you to model the distribution of gender in 10690 families with a random process, so that the distribution produced by this random process is nearly identical to the distribution of boys in the empirical data?

Figure 1. Geissler task (student handout translated from German)

When participants opened the TinkerPlots file, the screen looked like Figure 2.

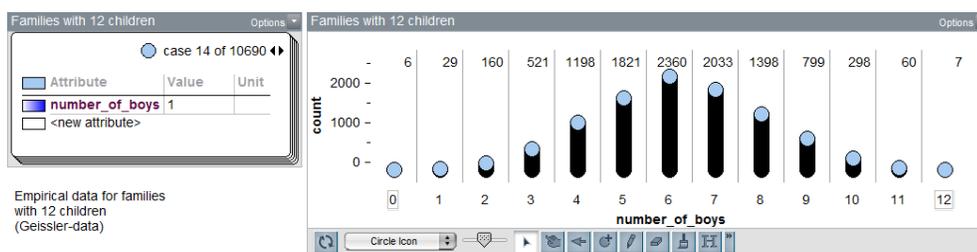


Figure 2. TinkerPlots screen at the beginning of students' work

4.4. PARTIAL ANALYSIS OF THE TASK

In preparation for the study we analyzed the challenges of the task from the perspective of participants' knowledge that they had ideally acquired in the MMDC course. This type of task analysis (as in Hadas & Hershkowitz, 2002) is supposed to show possible paths of learners solving the task and reveal obstacles and misconceptions of learners when working on the task. During this analysis we identified several potential student difficulties and prepared consecutive interventions for each part of the task. The interventions were related to expected stochastic difficulties and expected technical difficulties with TinkerPlots. Our goal was to be prepared for the most common difficulties (research question 3). We also wanted to see where they needed support and whether the prepared support was helpful to the students.

We expected the most challenges in parts three and part four of the task. Part three is about comparing original Geissler data with simulated data and part four is about adjusting the model set up in part two. For this reason we analyze both parts concerning what we expected "ideal" MMDC course participants to do. Then we identify possible difficulties within the parts and describe in more detail where we provided interventions when our students struggled.

In part one we expected the participants to interpret the Geissler data successfully. We did not expect many difficulties with this part, but in the case students could not describe the distribution we prepared some interventions to keep them working (see Appendix C).

Part two of the task included setting up a model in TinkerPlots that (re-)produced the data given in Geissler's dataset. At this stage an assumption about the probability of a "boy birth" was necessary. Because $p(\text{"boy"}) = 0.5$ is a common assumption based on everyday knowledge, at first we expected our participants to state that the probability for a boy and a girl was equal. The model underlying this assumption can be easily set up in TinkerPlots

Sampler with a spinner with two equal sections or with a mixer containing two balls. There are 10690 repeats for 10690 families and 12 draws for 12 children for each family are necessary (Figure 3). The participants learned to set up TinkerPlots Sampler like this in the course (for detailed interventions on part two see Appendix C).

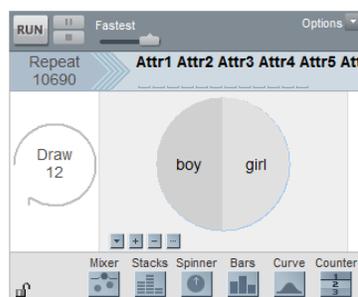


Figure 3. TinkerPlots Sampler with $p = 0.5$

The starting point for comparing Geissler data with simulated data could be a distribution emerging from the simulation with the Sampler in Figure 3 with $p(\text{boy}) = 0.5$, as shown in the bottom of Figure 4. The graph at the top of Figure 4 shows the original Geissler data.

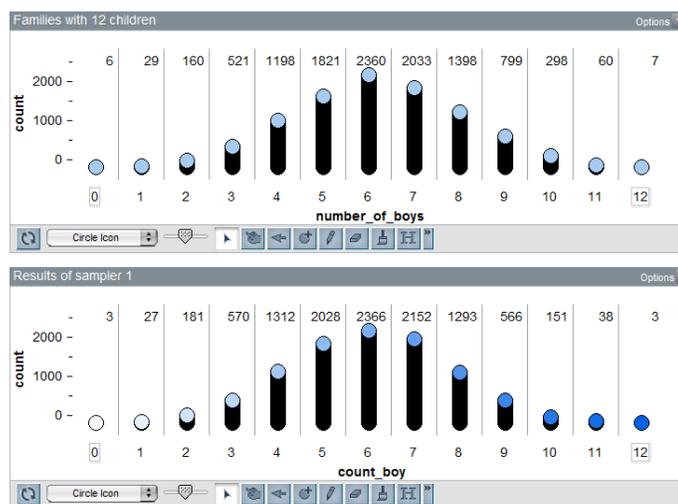


Figure 4. Distribution of Geissler data (top) and simulated data for $p = 0.5$, $n = 10690$ (bottom)

The setup shown in Figure 4 is optimal for doing a graphical comparison. It could of course occur that students will select a different TinkerPlots graph for displaying the simulated distribution or change the Geissler representation to a different graph. Students also may not exactly align the graphs as shown in Figure 4. Exact graphical comparison would then become visually more difficult.

We next asked: How could participants compare both distributions? An “ideal” participant from the MMDC course would have two options:

- Looking at global graphical features (symmetry, shape) and making an intuitive judgment about whether the distributions are “similar”
- Doing numerical analysis of the frequencies in each bin, looking at deviations
 - Checking whether the frequencies in corresponding bins are “more or less the same”
 - Detailed comparisons of the frequencies in each corresponding bin, looking at numerical differences
 - Judging the size of the differences from a statistical point of view

These options are not mutually exclusive; reasoning with more than one option was expected. We elaborate the options.

Using global features (symmetry, shape) and making an intuitive judgment whether the distributions are “similar” Shape, peak, and symmetry look very similar at a first sight. Both distributions are unimodal and symmetric around the peak at 6. We did not expect our students to depict graphical distributional differences. With a different tool that supports other graphical comparisons, other judgments would have been possible. For instance, the graph shown in Figure 5 would provide a better basis for graphical comparisons. With some experiences of using such graphs one could “see” that the blue distribution is slightly “shifted to the right.” In other words, up to 6 boys the Geissler data have less families than expected from $p = 0.5$, whereas from 8 boys onwards the Geissler data have more families with that number of boys than expected.

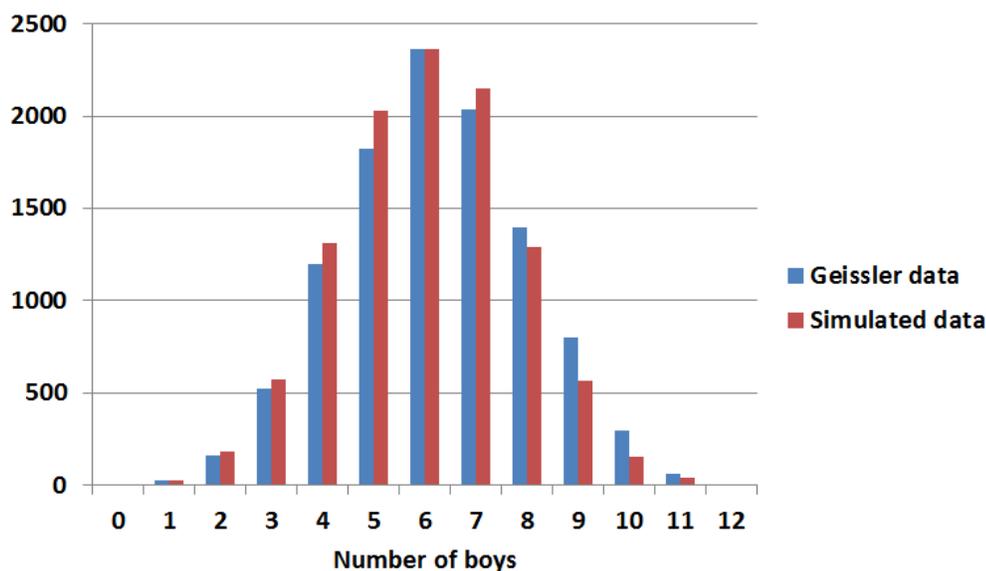


Figure 5. Geissler distribution (blue) and simulated distribution for $p = 0.5$ (red) in an Excel diagram

An expert modeler who would have noticed these differences between the bin frequencies could ask him/herself whether these differences between model and data seem to be within the usual random fluctuations to be expected or of a more substantial kind. We were interested in whether students would ask such kinds of questions or if they would be satisfied with the global similarity of the real data distribution and the simulated distribution.

Because we expected many students to be satisfied with the rough graphical comparison that is possible with TinkerPlots (as in Figure 4) and with their model at this stage, we prepared detailed interventions for this case. These interventions started with a question (See Appendix C) as intervention and ended with a hint to use the table (see Figure 6). This table was laying on the student work space from the beginning of the interview, but mostly disregarded until the prompt to use it was given. For an overview of all interventions for that situation see Appendix C: interventions 3_1_1 to 3_1_5. The intervention to use the table aimed at a bin-wise comparison of frequencies and a more substantial comparison, as in the next point.

Numerical analysis of the frequencies in each bin, looking at deviations A starting point may be the comparison of frequencies of cases in one pair of corresponding bins. One may identify huge differences between the distributions. For example, when comparing the frequency of the number of families with five boys in the Geissler distribution with the frequency in the simulated distribution, there is a difference of about 207 cases in bin 5. Differences may also be identified for several other bins. We did not expect that the bin-wise comparison would necessarily be done in a systematic way, so we prepared a table (Figure 6) for this situation. The table shown in Figure 6 allowed the students to document the real and the simulated data. This was meant to support the students' systematic analysis of the deviations between the Geissler data and the simulated data.

Number of boys in the families	Geissler distribution: Frequencies	Simulated distribution (p=0.5) Frequencies	Deviation (diff(Geissler; Simulated distribution))	Deviation upwards (+) or downwards (-)

Figure 6. Table prepared for comparing Geissler data with simulated data

In a systematic comparison using the filled-out version of the table shown in Figure 6, we expected that students might notice a pattern in the deviations similar to the pattern we identified in Figure 5.

The next question was: What sense can the students make of the observed differences with regard to the model chosen? We expected students to know that differences between simulated and real data are common, because they were taught that even different samples from the same model will differ from each other.

At this point we believed there were three ways the participants might take into account what they learned in MMDC:

- judging the differences between the bin frequencies referring to the $1/\sqrt{n}$ law,
- repeating the simulation (re-sampling) to get an idea of the size and pattern of variations between samples and then analyzing whether the real data fit well into this pattern (students might choose to concentrate just on the variation of one bin frequency and compare it to the Geissler frequency),
- making sense of the systematic pattern of + and - (see e.g., right column of Table 3) of deviations (which may suggest a higher probability for a boy birth might eliminate this pattern).

We start by outlining a possible way of arguing with the $1/\sqrt{n}$ law. The starting point here may be a judgment doubting that the size of the deviation can be attributed to chance variation. This can be judged more precisely with the $1/\sqrt{n}$ law. An obstacle might be that

the frequencies given in the Geissler data and in the simulated distribution are absolute frequencies. The preservice teachers were not taught how to judge differences in absolute frequencies, and therefore they had to calculate relative frequencies first and then use the $1/\sqrt{n}$ law. Another option would be to change the displayed empirical distribution to relative frequencies in the TinkerPlots plot, an option that was well-known to them (see Figure 7).

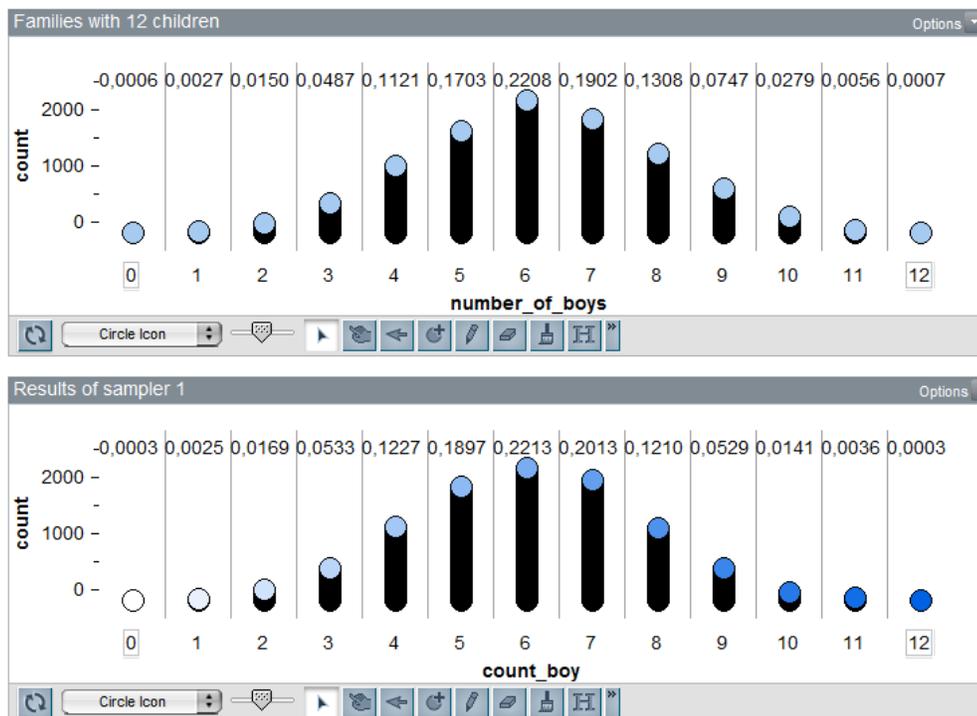


Figure 7. Distribution of Geissler data (top) and simulated data for $p = 0.5$ (bottom) with relative frequencies

Using these relative frequencies in combination with $n = 10690$ families (rounded $n \approx 10000$ and $\pm 1/\sqrt{n} = \pm 0.01$) they may look at differences above 0.01. Strictly speaking, we would have to distinguish the difference between the binomial model and the simulated data and the difference between the binomial model and the real data, so a double application of the $1/\sqrt{n}$ law is required, and respectively the 95% uncertainty has to increase when using the single $1/\sqrt{n}$ law. In the case of Figure 7, five bin frequencies differ as much or more than 0.01. Five out of twelve being outside the prediction intervals should lead to dissatisfaction with the model and therefore to the decision to improve the model if possible in the next part of the task.

In the event that students began reasoning with numerical differences between both distributions and stumbled at some point, we prepared interventions in advance. The bundle of interventions ranged from common questioning to the hint of using $1/\sqrt{n}$ law. For an overview of all interventions for that situation see Appendix C: Interventions 3_3_1 to 3_3_4.

A second way of reasoning is to resample. While continuously resampling and looking at some specific bins, the participants could get an intuitive feeling for the fluctuation of

the simulated data. Comparing this with the Geissler data and looking if the values occur in a certain range could lead to an informal way of rejecting the model.

The third way of reasoning is related to patterns and sizes of deviations. We thought a table might support this analysis. A table as such was not an unknown tool to the participants, but they may not have previously used tables for a systematic study of the differences between numbers in two columns. A possible use of the table from Figure 6 can be seen in Table 3. The occurring frequencies of both distributions are documented and the deviation is calculated (by hand). The last column shows the direction of the deviation.

Table 3. Comparison of the Geissler distribution and a simulated distribution with absolute frequencies in the table

Number of boys in the families	Geissler distribution	Simulated distribution ($p = 0.5$) in TinkerPlots	Deviation (diff(Geissler; Simulated distribution))	Deviation upwards (+) or downwards (-)
0	6	3	3	+
1	29	27	2	+
2	160	181	-21	-
3	521	570	-51	-
4	1198	1312	-114	-
5	1821	2028	-207	-
6	2360	2366	-6	-
7	2033	2152	-119	-
8	1398	1293	105	+
9	799	566	233	+
10	298	151	147	+
11	60	38	22	+
12	7	3	4	+

One aspect that could be noticed is the pattern in the deviations, and not only the deviation sizes. In Table 3 we see two positive deviations first, then only negative deviations and at the end again positive deviations. This looks quite regular and expresses a slight shift to the right of the Geissler data compared to the simulated data.

We faced representational limitations of TinkerPlots here: the program offers no easy way to track the differences of two distributions or to simplify the visual comparison of two distributions.

If the students recognize an obvious “nonrandom” pattern in the deviations as well as a shift between the distributions they should be unsatisfied with their model and choose to refine it in the next part of the task.

Here again we prepared several interventions for this way of reasoning. For an overview of these interventions see Appendix C: Interventions 3_4_1 to 3_4_4. The intervention bundle aimed to help students recognize a pattern in the deviations.

A goodness-of-fit test would be another possibility for the comparison, but this was not part of the MMDC-curriculum, so we did not take it into account in our analysis.

As a next step we expected our participants to improve their models in part four of the task. Finding a better model requires another look at the Geissler data to improve the first model.

We expected that the students would try out another probability for a boy’s birth and build a new model around a new estimate for this parameter. In principle they could question the implicit independence assumption of gender in successive births in a family as well, however we did not expect them to do so at this stage, for several reasons. First of

all, the independence assumption would be implicit in their models, as we expected them to choose their model by setting the number of draws to 12 (which implies independence) but not choosing a 12-step factory, where each step is independently combined with the next. So students may not be at all aware of having made an independence assumption, although this assumption was discussed at length in the course.

Looking for a different probability of boy birth may be the only option they would have. They may remember that they have heard or read that the probability for a boy is slightly higher than 50%. Or, they may have knowledge of how the change of p can affect the shape and location of a binomial distribution and use this knowledge to predict that a higher p for boys would graphically better fit the data than $p = 0.5$. The latter assumes that they have discovered the patterns in the residuals and can relate them to different binomial distributions.

To which probability can/should 0.5 be changed? A possible option for determining a new $p(\text{boy})$ is to determine the number of boys in the Geissler data and the total number of children and use the empirical proportion of boys in the sample as the estimated probability for the birth of a boy in those days. This proportion is 0.5168. This number cannot be calculated directly with TinkerPlots. The participants would have to calculate the proportion

$$\frac{29 \cdot 1 + 160 \cdot 2 + 521 \cdot 3 + 1198 \cdot 4 + 1821 \cdot 5 + 2360 \cdot 6 + 2033 \cdot 7 + 1398 \cdot 8 + 799 \cdot 9 + 298 \cdot 10 + 60 \cdot 11 + 7 \cdot 12}{12 \cdot 10690}$$

by hand (with a calculator). This $p = 0.5168$ can be set up in a model in an analogous way with a spinner in the TinkerPlots Sampler (Figure 8).

It would be necessary to judge again whether the new model is compatible with the Geissler data, or at least if it provides a better fit than the first model.

With this change, the differences in frequencies between the empirical and simulated distribution decrease and the data produced with this model do fit Geissler's data better. This can be seen while displaying the empirical and simulated data (Figure 9) and can be compared with the same ways of reasoning as before for model with $p(\text{boy}) = 0.5$.

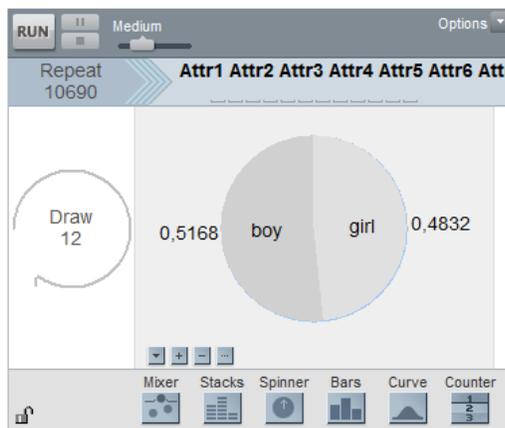


Figure 8. Refined TinkerPlots Sampler with $p = 0.5168$

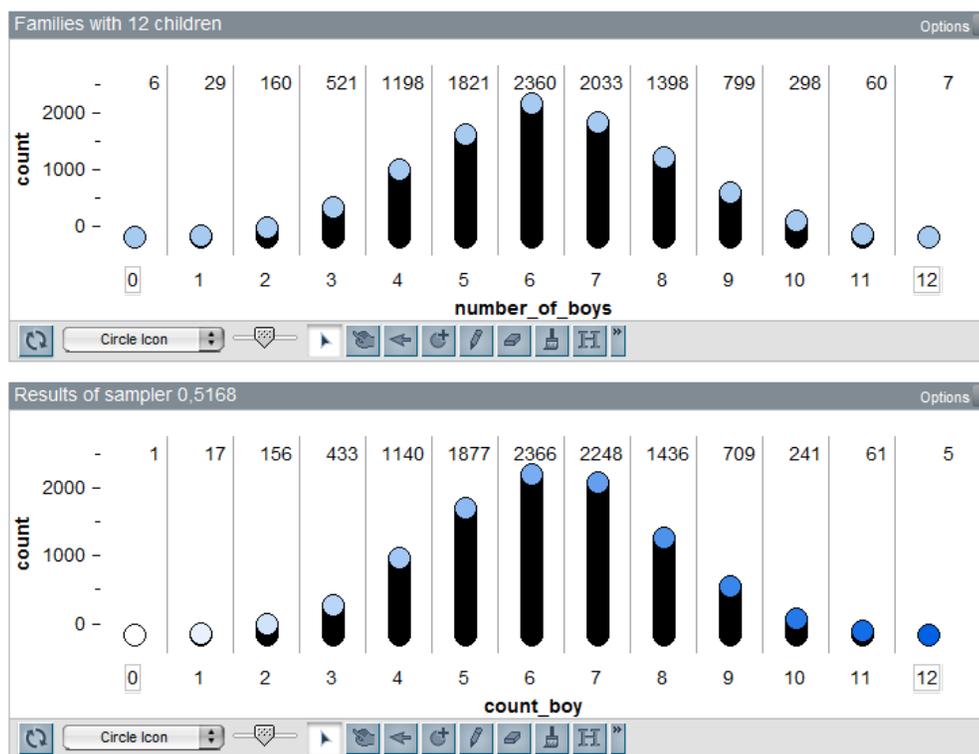


Figure 9. Distribution of Geissler data (top) and simulated data for $p = 0.5168$ (bottom)

Table 4 shows the deviations between the refined model and Geissler data and the direction of the deviations via + and – as before.

Table 4. Comparison of the Geissler distribution and the refined simulated distribution with absolute frequencies

No of boys	Geissler distribution	New Simulated distribution in TP	Deviation (diff(Geissler; New Simulated distribution))	Deviation upwards (+) or downwards (-)
0	6	1	5	+
1	29	17	12	+
2	160	156	4	+
3	521	433	88	+
4	1198	1140	58	+
5	1821	1877	-56	-
6	2360	2366	-6	-
7	2033	2248	-215	-
8	1398	1436	-38	-
9	799	709	90	+
10	298	241	57	+
11	60	61	-1	-
12	7	5	2	+

Table 4 shows the direction of the deviations between Geissler data and simulated data in its last column. With the previous model of $p = 0.5$ the deviations were downwards from two to seven boys and upwards for the remaining numbers of boys. With the new model of $p = 0.5168$ the deviations are a little more balanced.

With the more formal approach like the $1/\sqrt{n}$ law the students may change from absolute frequencies (see Table 4) to relative frequencies (see Table 5), using these relative frequencies in combination with 10690 families ($n \approx 10000$ and $\pm 1/\sqrt{n} = \pm 0.01$). In Table 5 only one frequency's absolute difference is larger than 0.01 (cf. Table 5, line with 7 boys).

Table 5. Comparison of Geissler distribution and refined simulated distribution with relative frequencies

No. of boys	Geissler distribution (relative frequencies)	New Simulated distribution in TP (relative frequencies)	Deviation (diff(Geissler; New Simulated distribution))	Deviation upwards (+) or downwards (-)
0	0.0006	0.0001	0.0005	+
1	0.0027	0.0017	0.0012	+
2	0.0150	0.0156	0.0004	+
3	0.0487	0.0433	0.0088	+
4	0.1121	0.1140	0.0058	+
5	0.1703	0.1877	-0.0056	-
6	0.2208	0.2366	-0.0006	-
7	0.1902	0.2248	-0.0215	-
8	0.1308	0.1436	-0.0038	-
9	0.0747	0.0709	0.0090	+
10	0.0279	0.0241	0.0057	+
11	0.0056	0.0061	-0.0001	-
12	0.0007	0.0005	0.0002	+

An expert modeler may notice that the tails of the simulated distribution are lower than in the Geissler distribution and the simulated distribution is higher in the middle (bins 5-8). Biehler (2005) explained:

School teaching could stop here and inform students that this deviation is well known by statisticians and biologists and gave rise to several attempts of explanation. An obvious hypothesis is that identical twins may be responsible for the surplus. This was checked but this did not completely account for the data. The independence assumption or the constant probability assumption has to be put into question. As probably most of the children that belonged to one family in the 19th century had the same mother and father, there may be acting some hidden biological factors inside families. (p. 27)

We were therefore prepared for some students to notice this new regularity but expected that they would, in general, be satisfied with their new model.

4.5. DATA COLLECTION

Approximately two months after the MMDC course concluded, 14 randomly chosen participants were invited to take part in the interview study. It was held in a laboratory setting at the University of Paderborn in September 2014. The participants were asked to work on the Geissler task in pairs (also assigned randomly) and were given an exercise sheet (Figure 1), and the Geissler dataset in a TinkerPlots file (Figure 2). We observed the seven pairs of participants as they took part in approximately 60-minute interviews. In the interviews, they were asked to think aloud as they solved the Geissler task in pairs. The

interviews were videotaped. The activities on the computer screen and the communication of the participants were recorded with Camtasia screen recording software. Additionally, we collected the TinkerPlots files and the notes the participants wrote on paper. The communication and the action in the software were transcribed.

23	B	Yes. Shall I take a Sampler?
24	A	Yes, that's fine.  <i>(TP: Sampler is generated.)</i>
25	B	Okay. Well, I would say, we have two possibilities: boy and girl, so we need two balls.

Figure 10. Example of transcript

These transcripts were the basis of our qualitative analysis. As shown in Figure 10, the transcripts contained the communication of the participants (here: B and A) and interviewer (I) and the action of the participants with TinkerPlots (in *italics*). Furthermore, TinkerPlots screenshots were incorporated in the transcripts for a better understanding.

4.6. INTERVIEW PROCEDURE

Because this kind of task was new to our students, we expected difficulties at several stages as mentioned in our task analysis above. All in all, after our task analysis, we identified several main obstacles for our participants:

- Participants do not find a starting/access point for the Geissler task
- Participants have technical difficulties with TinkerPlots (e.g., setting up the Sampler, plotting outcomes adequately)
- Participants have difficulties with the comparison of simulated data and Geissler data
- Participants do not know which aspect of their model can be changed to improve their model

To forestall this, we decided to support the participants by giving interventions (in the case that they would not proceed on their own) in the sense of the idea of minimal help (Leiss, 2007). That is why we prepared several interventions on different levels on the basis of our task analysis: feedback interventions, strategic interventions, and contextual interventions. These interventions were written down on paper for the interviewer, so that the interviewer could use the interventions in their exact wording. This paper was accessible to the interviewers only.

When one of the difficulties of the bulleted list (see above) occurred, we provided the participants with interventions in this order: feedback interventions, strategic interventions, and contextual interventions. At first we gave them feedback interventions; in the case that this was not enough to continue on their own, we provided them with stronger interventions (strategic interventions and contextual interventions). When the students were reasoning about the comparing subtasks and could not proceed on their own, we supported them with interventions in the direction of their line of argument. For example, if the students tried to reason with the $1/\sqrt{n}$ law but had difficulties remembering details of this law, the interviewer gave interventions prepared for reasoning with the $1/\sqrt{n}$ law. If the students argued with the $1/\sqrt{n}$ law afterwards correctly, no further interventions were given. For example, students using the $1/\sqrt{n}$ law for comparing both distributions were given no intervention on using the table for deviations (Figure 6). And vice versa, if students reasoned with deviations and received interventions for using the table to proceed, they got no intervention on using the $1/\sqrt{n}$ law.

Interventions were prepared and staged for all subtasks; but it was up to the interviewer to recognize the students' problems or ways of reasoning and to choose the fitting intervention for these situations. Because of the graduation of the interventions, the interviewer could select the appropriate intervention by listening to the argumentation of the students.

These interventions were intended to help participants continue the modeling process on their own. Table 6 describes the interventions (the types in column 1 are adapted from Leiss (2007); we added our interpretations) in a general form with additional examples chosen from our data.

Table 6. Intervention types (adapted from Leiss, 2007)

Kind of intervention	Description	Example
Feedback	Feedback interventions give feedback in the sense of whether the procedure is done rightly or wrongly (but without giving hints to specific content).	"This is not done correctly."
General strategic	General strategic interventions try to influence the solving process in a positive way without giving mathematical help directly.	"Have a look on the simulation scheme. Maybe it can help you."
Contextual strategic	Contextual strategic interventions include hints that could be used directly for solving the problem. They are stronger (and have a closer relation to the task) than general strategic interventions.	"Try to write down the deviations of the number of boys in both distributions systematically."
Contextual	Contextual interventions rely on the context of the task directly. They give a specific advice for solving the problem.	"You need to set up the variable Count_boys."

5. METHODS AND LEVELS OF DATA ANALYSIS

We conducted a three-level-analysis. At first we wanted to know whether the pairs were capable of solving the Geissler task in general (see research question 1). In level 2 and 3 we wanted to have a closer look in the detailed solving processes of our pairs. In level 2 we want to identify which phases occurred in the modeling process and during which phases interventions were necessary (research question 2 and 3). Level 3 analysis was done to gain insight about participants' cognitive processes and their argumentation when comparing the empirical and simulated distributions when working on the Geissler task (research question 4).

5.1. ANALYSIS LEVEL 1

According to the four subtasks on the exercise sheet (Figure 1) we identify four steps for solving the Geissler task successfully.

1. Subtask 1 is solved successfully if the Geissler distribution is described by shape, center, and symmetry, informally.
2. Subtask 2 is solved successfully if the model set up to produce Geissler data is displayed in a TinkerPlots Sampler with $p(\text{"boy"}) = 0.51$, $\text{draw} = 12$, and $\text{repeat} = 10690$.
3. Subtask 3 is solved successfully if the Sampler is run at least once, the random variable "number of boys" is defined correctly, the distribution of the random variable "number of boys" is plotted in TinkerPlots, and the Geissler distribution and the simulated distribution are compared correctly via one of the following approaches:
 - a. Looking at global features (symmetry, shape) and making an intuitive judgment about whether the distributions are "similar."
 - b. Documenting the frequencies in each bin, looking at deviations (differences between the bin frequency of simulated data and of Geissler data).
 - c. Using the $1/\sqrt{n}$ law, respectively the \sqrt{n} law, to evaluate the size of the deviations in the previous approach.
 - d. Using a resampling approach along with one of the approaches above.
4. Subtask 4 is solved successfully if the participants judge their models of task 2 as fitting or not fitting to Geissler data (with respect to approaches of subtask 3).

Coding was done on the work of each pair on the tasks. The whole process of solving the Geissler task was taken into account for analysis level 1. We coded each pair's work a "yes," when a Geissler subtask was solved successfully. Analysis level one was a simple report of whether or not students could solve the Geissler tasks successfully. More details of students' solving processes are shown in analysis levels 2 and 3.

5.2. ANALYSIS LEVEL 2

The goal of analysis level 2 was to identify structures in the solving processes of the participants. We wanted to determine which phases appeared and in which ways and when the participants had to be supported by interventions. To gain insight into the occurrence and order of the phases of a modeling cycle (research question 2), we used qualitative content analysis (Mayring, 2015) to reveal the sequence of processes. Additionally, process

diagrams were constructed to illustrate structures in the pairs' work and enable us to compare the processes of all pairs. These process diagrams also showed typical behaviors of our preservice teachers when working on such a task.

The qualitative content analysis conducted had the goal “to filter out a particular structure from the material...The text [transcript] can be structured according to content, form and scaling” (Kohlbacher, 2006, p. 12). This method of analysis is especially useful when analyzing a large amount of data and when searching for structures (like solving process paths) in the data. A further advantage of qualitative content analysis is the “systematic, rule-bound procedure,” the “categories as the focus of analysis,” and the “theory-guided character of the analysis” (Mayring, 2015, pp. 369-371). To set up a category system with exact definitions and key examples is inevitable when working with qualitative content analysis. The exact ordering of qualitative content analysis is given by the sequence (see Mayring, 2015) displayed below.

1. Identification of object of research
2. Formulation of selection criteria
3. Generation of categories (deductively and inductively)
4. Coding of selection of cases
5. Modification of categories
6. Constructing coding manual with definitions and key examples
7. Coding the whole material
8. Frequency analysis of occurrence of codings

For our purpose a structuring qualitative content analysis seemed to be adequate, because we had a large number of transcripts and we wanted to structure and evaluate the phases of our participants when working on the Geissler task. The main goal during analysis was to structure the transcripts on the basis of phases (Tables 7 and 8). We also wanted to develop diagrams for each pair to depict the processes they followed and to compare the processes of all pairs to one another.

Next, we addressed research question 2: “Which structure and phases can be identified when preservice teachers are working on the Geissler task with TinkerPlots?” and research question 3: “In which circumstances do the preservice teachers need support?” We analyzed all transcripts and Camtasia recordings for our seven pairs. One coding unit was a unit of meaning. A unit of meaning could be the uninterrupted talk of a student and/or related activities in the software. Coding units were assigned to the codes disjointly (no multiple coding). Coding was done on the transcript and the video recordings by the second and third author. Coding disagreements were discussed by the second and third author until an agreement was reached.

On the basis of our expectations of possible phases to solve the Geissler task, we deductively distinguished between phases that belonged to statistical reasoning (statistical phases, see Table 7) and phases that covered action with TinkerPlots (TinkerPlots phases, see Table 8). Both phases (Statistical phases and TinkerPlots phases) were disjoint subsets of our transcripts. TinkerPlots activities written in italics in the transcripts were transcribed as TinkerPlots phases and occurred within statistical phases (see Table 7).

Table 7. Statistical phases

Statistical phases	Description
<i>Phase D</i>	<i>Describing and interpreting the given Geissler data</i>
<i>Phase G</i>	<i>Generating a model to reproduce data</i>
<i>Phase A</i>	<i>Analyzing Sampler results in TinkerPlots</i>
<i>Phase C</i>	<i>Comparing the sampled distribution with the Geissler distribution</i>
<i>Phase V</i>	<i>Validating the model</i>

After ending phase V, another run through the statistical phases may be appropriate if the model does not fit the Geissler data. Exact definitions and key examples of the statistical phases can be found in Appendix A.

Table 8. TinkerPlots phases

TinkerPlots phases	Description
<i>TP phase S</i>	<i>Setting up the Sampler</i>
<i>TP phase I</i>	<i>Identifying the random variable: defining the result attribute “number of boys”</i>
<i>TP phase P</i>	<i>Plotting the result attribute</i>
<i>TP phase R</i>	<i>TinkerPlots Repeat</i>
<i>TP phase E</i>	<i>TinkerPlots Exploration (working in TinkerPlots without moving to another phase)</i>

Exact definitions and key examples of the TinkerPlots phases can be found in Appendix B. As we mentioned in subsection 4.5, we supported the participants when they were not able to proceed. All types of interventions, their exact definitions, and key examples can be found in Table 6.

Because we had a large number of transcripts, we used MAXQDA, a qualitative data analysis tool for computers (Kuckartz, 2012, p. 132). MAXQDA offers the possibility of structuring the transcripts according to several codings and supports several analysis methods that enable the interpretation of the given data. Figure 11 shows the elements of MAXQDA. On the right side the transcript can be displayed (in connection with video and audio file), on the left side codings can be generated in several ways (deductive, inductive, in-vivo). Passages of the transcript can then be marked and allocated to the several codes on the left side.

The screenshot displays the MAXQDA software interface. On the left, there is a 'Liste der Dokumente' (List of Documents) pane showing a tree view of documents and a 'Codesystem' (Code System) pane showing a hierarchical list of codes. The main window shows a transcript with a list of utterances, each with a time stamp and a speaker label (I, C, S). The transcript text includes:

- 01 25 I Ja, sehe ich auch so. 00:02:43-9#
- 04 26 C Okay, Nummer Zwei. (13.0) Also wir sollen das jetzt hier oben versuchen mit der Zufallsmaschine, dass das Ähnliche rauskommt, oder am besten dasselbe, ne. Ok. 00:03:07-0#
- 07 27 S Dann würde ich hier auch schon mal für die Wahrscheinlichkeit, also p wäre 0,5, kann man ja von ausgehen. *Das Fenster der Zufallsmaschine wird erstellt.*
- 09 28 S Und jetzt zwei Merkmale mit J und M für Junge und Mädchen. *TP: In den Kreis a wird ein „J“ und in den Kreis b ein „M“ eingetragen.*
- 14 29 C Ja. 00:03:24-7#
- 16 30 S Dann Ziehungen zwölf,ne. Für zwölf Geburten, also für die zwölf Kinder. 00:03:39-3#
- 18 31 C Achso, mhm (3.0) Echt? 00:03:46-0#
- 20 32 I Macht erst mal, also wir helfen euch weiter wenn ihr gar nicht mehr weiter wisst. Probiert erst mal aus.
- 22 33 C Also ich muss erst mal überlegen. Es gibt wie viel Familien? 00:03:54-5#
- 24 34 S 10690 Familien. 00:03:58-7#
- 26 35 C So und jede Familie hat zwölf Kinder. (2.0) Dann müssen wir doch bei Durchgänge die Familie machen oder nicht? Für jede Familie ein / 00:04:09-6#
- 28 36 S Ja. *TP: Bei Durchgänge wird die Zahl 10690 eingetragen.*

Figure 11. MAXQDA screen

MAXQDA offers several analysis tools such as frequency tables of codings, contingency tables of codings, etc. A powerful tool in MAXQDA is the document portrait. This allowed us to represent the problem-solving structures of the pairs of participants and compare them. The document portrait always displays an array of 30 x 40 colored tiles, independent of the length of the transcript (Figure 12).

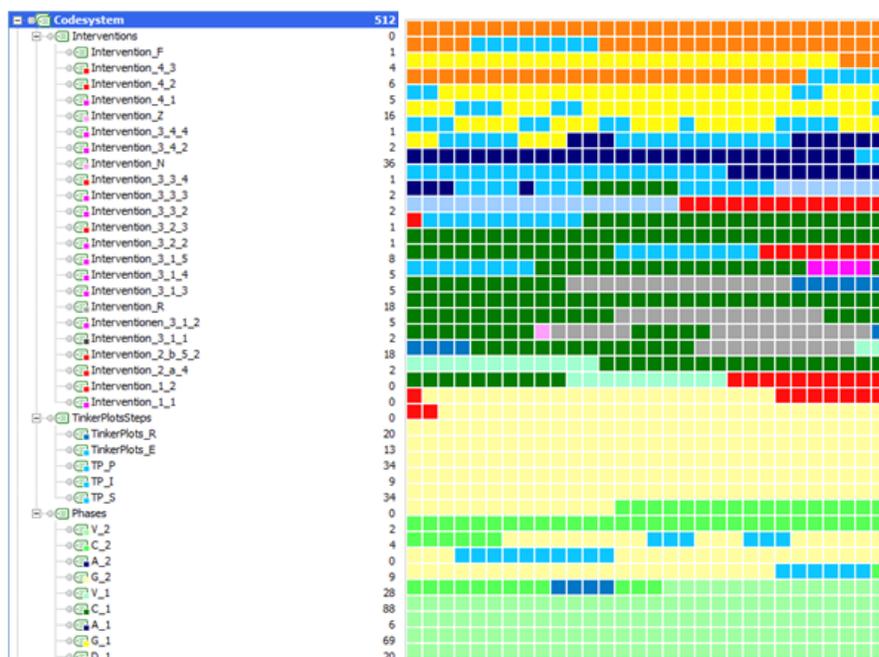


Figure 12. Sample MAXQDA document portrait with related codings

Table 9 lists all colors given to codes shown in the document portrait in Figure 12. Not every coding got its own color, because this would not be distinguishable in the document portrait. For example all interventions were coded separately, but all got shades of red as their color.

Table 9. Colors in document portrait related to phases

Color	Codings related to phase
Orange	Phase D: Describing Geissler data
Shades of blue	Different TP phases
Yellow	Phase G: Generating a model
Dark green + grass green	Phase C: Comparing distributions
Light green	Phase V: Validating model
Red, pink and purple	Interventions, prepared
Grey	Intervention as inquiry

Every tile refers to a code assigned to the transcript. The arrangement is chronologic by appearance in the transcript and shows by color the different codings. The length of each coding is taken into account via the length of tiles of the same color. The longer a sequence in the portrait, the longer it is in the transcript. This “length” is not proportional to the time of the episode. Interpretations have to take this into account.

To provide a better impression of a document portrait, we describe an exemplary document portrait produced by MAXQDA in the results section for analysis of level 2 (Figure 13). The document portraits were created to help us to find structure and to compare the processes of all pairs.

5.3. ANALYSIS LEVEL 3

At this analysis level we analyzed which approaches the participants used to compare the Geissler distribution to their simulated distributions and how they evaluated their models. To address the research question 4: “How do the preservice teachers reason about their models in the face of real data?,” we took our process structure (see analysis level 2) and looked more closely at the communication and TinkerPlots actions of participants in several phases. We focused, in particular, on the codings related to the “compare” and “evaluate” phases (see analysis level 2) in the process. These episodes were chosen because they were “crucial episodes” (Voigt, 1984) and they were analyzed with interpretive methods. The selection of these crucial episodes was done with regard to the research questions. The interpretive methods differed from the content analysis methods, which were used in analysis levels 1 and 2 because for analysis level 3 a turn-by-turn analysis of the crucial episodes in the transcripts was conducted (Krummheuer & Naujok, 1999).

6. RESULTS

We first describe the results of each analysis level and then summarize and discuss our findings.

6.1. RESULTS OF ANALYSIS LEVEL 1

As shown in the right column in Table 10 all pairs successfully completed subtasks 1-4 (at times, with interventions, reported later). Differences were only observed on the reasoning for comparing the Geissler distribution with the simulated distribution. In addition, three of the seven teams compared the Geissler distribution and the simulated distribution via frequencies (documented in Table 4), and three other teams used the $1/\sqrt{n}$ law to identify differences between Geissler data and simulated data. None of the teams showed reasoning with resampling.

Table 10. Results of analysis level 1

Task	Component	Team B&B	Team R&U	Team B&S	Team D&S	Team S&S	Team F&M	Team D&P	Total
1	Shape/Symmetry	+	+	+	+	+	+	+	7
2	$p = 0.5168$	+	+	+	+	+	+	+	7
2	Draw = 12	+	+	+	+	+	+	+	7
2	Repeat = 10690	+	+	+	+	+	+	+	7
3	Sampler run at least once	+	+	+	+	+	+	+	7
3	Number of boys correctly	+	+	+	+	+	+	+	7
3	Distribution is plotted	+	+	+	+	+	+	+	7
C	Global feature	+	+	+	+	+	+	+	7
C	Frequencies documented and deviations calculated	+	-	-	+	-	-	+	3
C	$1/\sqrt{n}$ law	-	+	-	-	+	+	+	4
C	Resampling	-*	+	-*	-*	+	+	+	(4)*

4	Closing reasoning about the model	+	+	+	+	+	+	+	7
---	-----------------------------------	---	---	---	---	---	---	---	---

Note: “+” means successfully used; “-” means not successfully used or not used at all; “_*” or “+*” means used with repeated simulations

All pairs marked with * ran their Sampler more than two times, but none of them constructed arguments with aspects of this repetition. It was our sense that this was not a “resampling” approach.

We can state that all participants were able to solve the Geissler task successfully with interventions. This is shown in Table 10 under subtasks 1 to 4. Each subtask was solved by every pair correctly. Differences occurred in the ways participants reasoned about task 3. We will have a closer look at specific problems that occurred in section 6.2.

6.2. RESULTS OF ANALYSIS LEVEL 2

In analysis level 2 we investigated two research questions: “Which structure and sequence of phases can be identified when preservice teachers are working on the Geissler task with TinkerPlots?” and “In which circumstances do the preservice teachers need support?”

We describe the case of Francis and Marc (F&M) first. In this example we also explain our procedure for analyzing our data via document portraits.

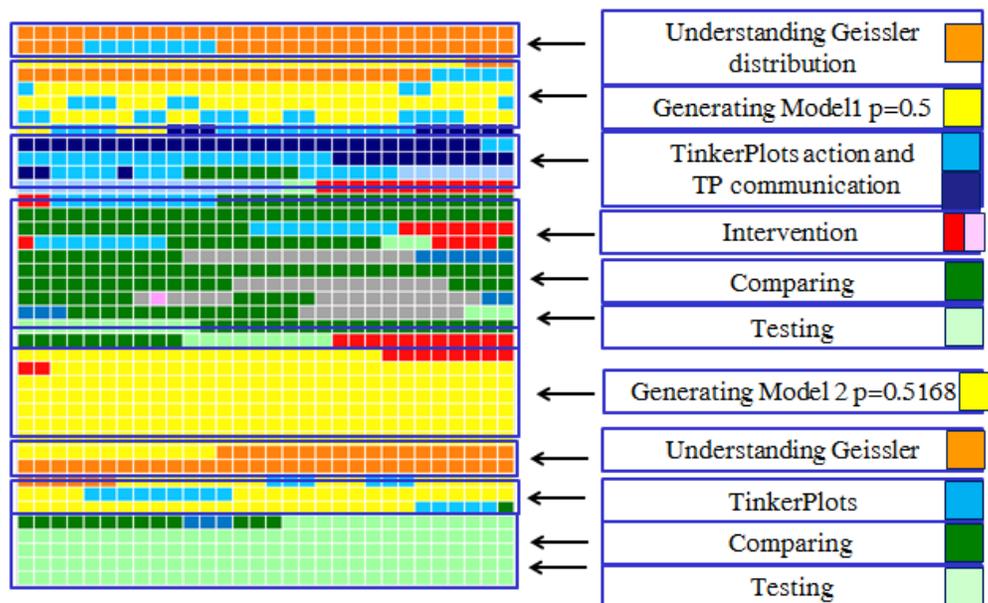


Figure 13. The MAXQDA document portrait of Francis and Marc and its structure

The document portrait (Figure 13) shows the whole process Francis and Marc used when working on the Geissler task. At first they described and interpreted the Geissler distribution (orange). Then they continued with task 2 (phase G1: generating a model), which is displayed as yellow tiles in the document in combination with bright blue tiles (TinkerPlots activity, phase TP_S: setting up the Sampler). The dark blue tiles that follow

represent the process of running the Sampler and plotting the distribution of the variable “number of boys” (phase TP_I). Then the comparison phase (green) follows; here Francis and Marc compared the Geissler distribution with their simulated distribution based on their model (phase C1). There are red and pink tiles between the “green” process tiles. These are interventions done by the interviewers (pink tiles show strategic, red tiles contextual, interventions). The grey tiles display inquiry questions by the interviewer, but no interventions. After a short phase of validating the model (light green phase, V1) and two interventions by the interviewers, Francis and Marc revised their model (light yellow phase, G2). This phase turned out to be long (in comparison to the other phases) because of the process of calculating the probability (“boy”) from the Geissler data. After comparing the Geissler distribution and the new simulated distribution based on the revised model (bright green phase, C2), they finally validated their revised model (light green phase, V2), and responded to the main question within this phase (“Does the TinkerPlots Sampler offer the possibility of modeling the distribution of gender in 10690 families with a random process, that the distribution produced by this random process is nearly identical to the distribution of boys in the empirical data?”).

We produced document portraits in MAXQDA for every team. Looking at these document portraits we can identify two different approaches used by our participants to solve the Geissler task. Five of the seven pairs (all pairs except for Barbara & Steffi and Denise & Paula) had approaches consisting of the following steps:

1. Understanding the Geissler distribution
2. Generating a model with $p = 0.5$
3. Setting up the model in TinkerPlots
4. Comparing Geissler and simulated distribution
5. Testing and rejecting the first model
6. Generating model with $p = 0.5168$
7. Comparing again
8. Testing again

Beyond that, two pairs (Barbara and Steffi, Denise and Paula) ran a third cycle. For both of these pairs we identified a different approach consisting of the following steps.

1. Understanding the Geissler distribution
2. Generating model 1 with $p = 0.5$
3. Setting up the model in TinkerPlots
4. Comparing Geissler and simulated distribution
5. Testing and rejecting the first model
6. Generating model 2
7. Setting up model 2 in TinkerPlots (with $p = 0.54$ by Barbara & Steffi, just guessing another p for the model with no reasoning; and $p = 0.5$ but repeat = 1000 for Denise & Paula)
8. Comparing Geissler and new simulated distribution
9. Testing and rejecting the second model (both models were refused immediately after plotting the results because of their obvious deviations from the Geissler data)
10. Generating model 3 with $p = 0.5168$
11. Setting up model 3 in TinkerPlots
12. Comparing Geissler and new simulated distribution
13. Testing third model

All in all we can state—relating to research question 2—that a general cycle of *generating a model, setting up the model in TinkerPlots, comparing results with the Geissler distribution, testing the model*, is visible. This cycle was run twice (5 of 7 pairs) or three times (2 of 7 pairs).

We now look at the second aspect of analysis level 2: *Structuring the process – Analysis of structures and necessary interventions in the modeling phases* and research question 3: “In which circumstances do the preservice teachers need interventions?” In Table 11 we see the distribution of interventions separated by statistical phases (cf. Table 7).

Table 11. Overview of interventions

Pair	Phase D	Phase G	Phase A	Phase C	Phase V	In total
<i>Brooke and Bella</i>	0	0	0	4	3	7
<i>Rosi and Ulla</i>	0	2	0	8	1	11
<i>Barbara and Steffi</i>	0	4	0	3	1	8
<i>Dennis and Sandra</i>	0	5	0	9	0	14
<i>Sean and Silke</i>	0	4	0	4	6	14
<i>Francis and Marc</i>	0	2	0	1	2	5
<i>Denise and Paula</i>	0	4	0	6	2	12
<i>In total</i>	0	21	0	35	15	71

In all, 71 interventions were needed for the seven pairs. Geissler task phase D (understanding the Geissler distribution) was solved successfully without interventions. Table 11 and all document portraits indicate that nearly half of the interventions occurred in phase C, so the comparison of the distributions seemed to be the most difficult aspect for participants. TinkerPlots was handled well by the participants; mainly, only a problem with bin width occurred (Intervention 2_b_5_2, TinkerPlots Phase P, included in Table 11 in Phase C) because the participants had problems setting the bin width to 1. The comparison part took a great deal of the working time on Geissler task. Also, refining the model and calculating $p(\text{boy})$ took much time.

To summarize the results of research question 3, we can identify several aspects of the document portraits. One point is that understanding the distribution (Geissler task 1, Phase D) and analyzing Sampler results (Phase A) were done without interventions, but in all other subtasks the participants needed interventions by the interviewer. The first TinkerPlots model (with $p = 0.5$) was also done without the interventions. The first (contextual) intervention was necessary in Geissler subtask 3 (comparing). Overall, we can say that interventions mainly occurred in the following phases: TinkerPlots Phase P: Plotting results (especially plotting the distribution in TP (Intervention 2_b_5_2)), Statistical Phase C: Comparing, Statistical Phase G: Generating a revised model, and Statistical Phase V: Validating the model.

From a teaching point of view, we can say that the prepared interventions proved to be adequate for helping participants solve the task.

6.3. RESULTS OF ANALYSIS LEVEL 3

In our level 3 analysis we focused on how the participants evaluated their models and which formal and informal reasoning could be identified. As stated in section 6.1, all seven pairs used global features, three of seven pairs used frequencies (documentation of frequencies and calculation of deviations), and four of seven pairs applied the $1/\sqrt{n}$ law.

All pairs started with a model with an equal probability ($p = 0.5$) for boys and girls. The other Sampler options, such as drawing twelve for twelve children in each family and

repeating 10690 times to represent all families in the dataset, were set up correctly by the participants. The crucial point for setting up the model seemed to be choosing the probability p . In this phase, however, the context did not play a fundamental role for the participants. No pair talked about the assumed probability of a boy's birth or the situation in Prussia in 1880. All pairs started without discussing the modeling phase with p ("boy") = 0.5. One major part of the task was comparing Geissler data to simulated data, because this was the only possibility for the participants to evaluate their models. All pairs started the first comparison of the Geissler data and the simulated data concentrating on global features. They compared shape, center, and symmetry only and stated their satisfaction with the model. At this stage interventions were necessary. After the first comparison phase, all pairs decided to go back to analyse the Geissler data again to have a better basis for revising their first model. For most pairs this was the first serious view of Geissler data. In the comparing phase, three pairs used the $1/\sqrt{n}$ law for comparing the Geissler distribution with the simulated distribution as a more formal way of reasoning. One pair used only global features for the comparison and no intervention could change this. The other three pairs used deviations and documented their directions in a table as a more informal way of reasoning. Most pairs tried several approaches, but reasoned only with one. Every pair resampled at some stages, but did not use the resampling as a line of argument. They resampled to get a feeling for the fluctuation of the data of one special event, for example, the frequency of five boys. All in all, the participants were able to cope with this new task and they were able to use TinkerPlots to express their model.

For deeper insight into participants' concrete argumentations, we describe the argumentation of Francis and Marc. We focus in detail on the following crucial episodes (see the following transcripts excerpts 1-4):

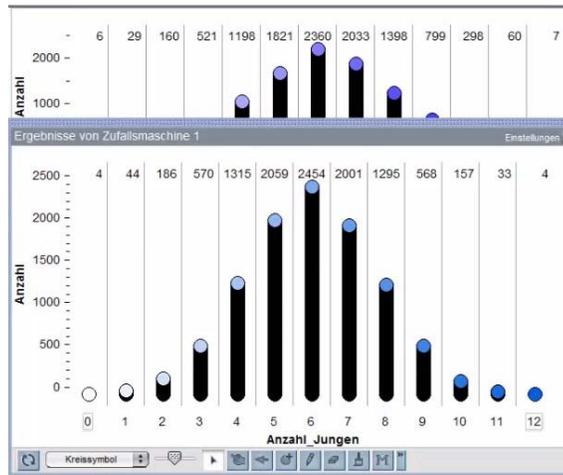
- Comparing the Geissler distribution with the simulated distribution ($p = 0.5$)
- Comparing the Geissler distribution with the simulated distribution ($p = 0.5$),
– after intervention
- Refining the original model, generating a new model
- Comparing the Geissler distribution with the new simulated distribution ($p = 0.5145$).

Francis and Marc showed good statistical reasoning in the whole course and worked on Geissler subtasks 1 and 2 successfully and did not need any intervention when working on them. In Geissler task part 1, Francis and Marc observed that the Geissler distribution looked "like a normal distribution," probably meaning a unimodal symmetric distribution with the highest probability in the middle and decreasing probabilities towards both sides. The normal distribution was not part of the MMDC curriculum. They might have encountered the Gaussian distribution in other contexts or just called unimodal symmetric distributions "normal." Furthermore, they mentioned that the distribution was not exactly symmetric, because in bin 7 there were more cases than in bin 5, although they emphasized that "it should be equal from the mathematical point of view."

In Geissler task part 2, Francis and Marc were convinced that due to the fact that there are two possibilities ("boy" and "girl") the probability of a boy or a girl birth is one-half. They set up this 50:50 model in the Sampler in TinkerPlots. Then they started the simulation and plotted the outcomes of the Sampler ($p = 0.5$, draw = 12, repeat = 10690) and began with Geissler task 3 (comparison). Here the transcript excerpt (see below) starts. M stands for Marc, F stands for Francis, and I stands for Interviewer in the transcript.

64 M So let's continue...(reading:) Compare the distribution of the simulation with the [Geissler] distribution. What do you recognize?

65 F



Yes, for most bars, it fits approximately. Of course it is not equal. It is possible that there are larger deviations, for example for “11”, there are 60 [cases] above [in the Geissler distribution] and now only 33 [cases], it is just the half! In bin 10, the amount of cases is also halved. So there are deviations, which may be caused by the variation due to random effects. Generally we see—as we see above—that it [the simulated distribution] is a normal distribution.

66 M

Even there are only half of cases in bin 11. This doesn’t matter... because 30, I mean 30 compared to 10690 is nothing, it is less than 1%. We can neglect it! The important question is whether the bars correspond approximately and whether they correspond based on their shape. We have again the normal distribution, and always two bars, which correspond to each other. There is a symmetry axis at bar 6.

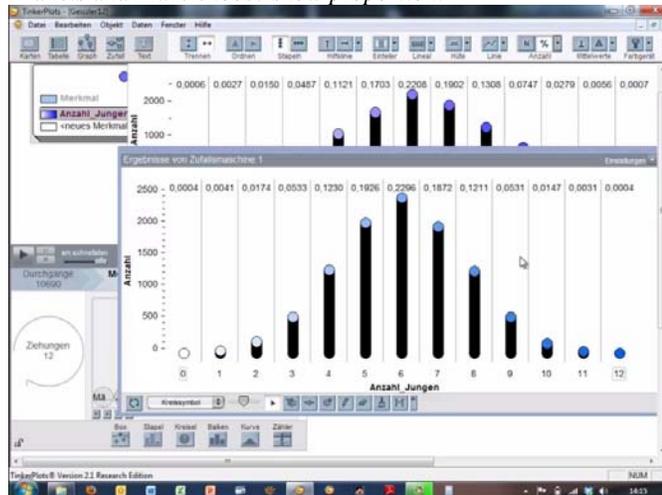
67 F

If we change the view to percentages we will see that they are very similar.

68 I

They click on % and disable N (absolute frequency)
 Maybe we should also choose *show proportion*, then we have two digits more.

Francis and Marc choose show proportion.



69 M

But now...it is a little bit confusing [for reading off the values].

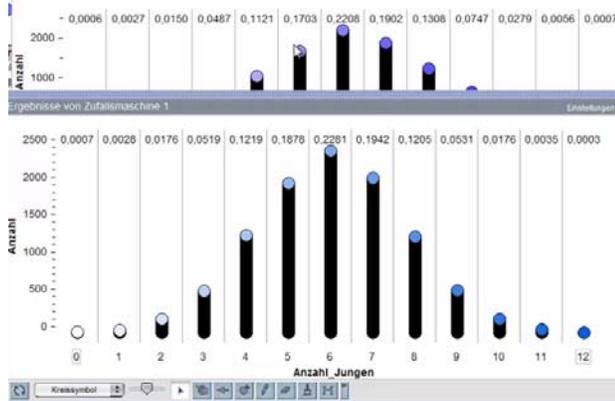
70 I

But it is more exact. You might enlarge the window, so you might better read that.

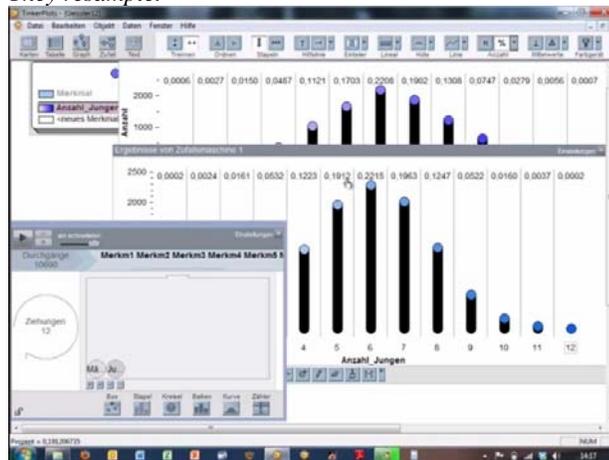
- 71 M *They enlarge the windows.*
Oh. So we see, that the bin of “6 boys” has 22% in both distributions. Well, rounded on the one hand 23%, on the other hand 22%. And here [bin of “7 boys”] we have in both distributions 19%, so we can say that they correspond.
- 72 I What about the bin of “5 boys”? What if you have a look at frequencies there?

This is a typical example of reasoning at the beginning of the comparison phase, which we also identified in the solution processes of the other pairs. The distribution of the simulated data had the same shape as the Geissler data. The participants knew from the course that results of a simulation fluctuate. Francis and Marc stated this (line 65) and seemed to have a first intuitive idea of rating the deviations in line 66. Marc seemed to be satisfied with their model. Support for this first assumption is that he said in line 66: “We have again the normal distribution, and always two bars which correspond to each other,” meaning that two bars have nearly the same height, like bar 5 and bar 7, bar 4 and bar 8, and so on. This seems to be close enough for him to the Geissler data and is a typical reaction we observed from many participants. This is similar to the way Konold and Kazak (2008) presented the idea of model-fit only via shape to younger students. A look at the shape of both distributions is the first approach to the comparison. This is an obvious start for the subtask and they judge the model as fitting. At this point interventions had been prepared. In the case of Francis and Marc, an intervention was not yet needed, because Francis wanted a closer look and changed the view from absolute to relative frequencies and rounded it verbally via a numerical-percentage strategy (line 71). However, they rounded percentages without examining the size of random fluctuations. This was a good approach to relativize the obtained absolute numbers. As most of the pairs, Francis and Marc only looked at one or two pairs of bins to compare the distributions. After a short interval the interviewer intervened to help the pair find some differences (see line 72), because the deviations were largest at bin 5. Here the next transcript excerpt continues (see below) and shows how Francis and Marc compared the distributions after the intervention.

- 73 M Okay, we have a deviation there, 2% [in bin 5].
- 74 F If we repeat the whole stuff, it would be possible to have less than 17%.
- 75- *Students are a little bit confused about the technical aspects of*
79 *TinkerPlots while resampling.*
They repeat the simulation; the new simulated distribution is shown in
the lower plot.



- 80 M Oh...okay. We see [they look at the bin of 5 boys]... We have not 17% anymore, but 18.7%. We could round it to 19%. We have a little bit more than before, but the values will vary by only a few, so that we can compare them to the empirical data from Prussia in 1880.
- 81 F The law of large numbers will help us to decide in which way the values will level.
- 82 M Yes, this was the formula with... anything with $1/\dots$
- 83 F Squareroot n.
- 84 M Squareroot n.
- 85 I Okay.
- 86 M Minus Squareroot n? Something like this.
- 87 I $1/\text{Squareroot } n$. What would it mean in our case?
- 88 M Plus Minus $1/\text{Squareroot } n$. That's it!
- 89 F So, n is larger than 10000, precisely 10690. So let's calculate with 10000. The squareroot of 10000 is 100, so that means that 95% of the cases are in the interval of plus/minus 1%.
- 90 I That's correct. And? Does it fit here...in our case?
- 91 M No!
- 92 I *(laughing)* Why no?
- 93 F This is more than 1% at [bin] four and [bin] five.
- 94 I You have already repeated the simulation. Do you want to repeat it once again? Before we had 3%, didn't we?
- 95 F Yes, this was more.
They resample.



- 96 M So we have a width of the probability we have already had, ranging from 19.4 to 17.8. This is percent [*they compare the frequencies of bin of 5 boys in both distributions*].
- 97 I Yes, this would not correspond with your findings above. At bin 8 boys it also does not fit with the 1% [*they compare the frequencies of 8 boys in both distributions*].

With a look at bin 5, Marc recognized a deviation of 2% between the empirical and simulated data in families with five boys. Hereupon, Marc proposed repeating the simulation again (line 74: “If we repeat the whole stuff, it would be possible to have less than 17%”). Perhaps he was thinking that this was an outlier and the simulated result at bin 5 would fit better to the empirical data after resampling. After resampling, Marc was satisfied and rounded again with a numerical-percentage strategy (line 80), but Francis now wanted to rate the deviations. Again, he was the one who wanted to have more details. He

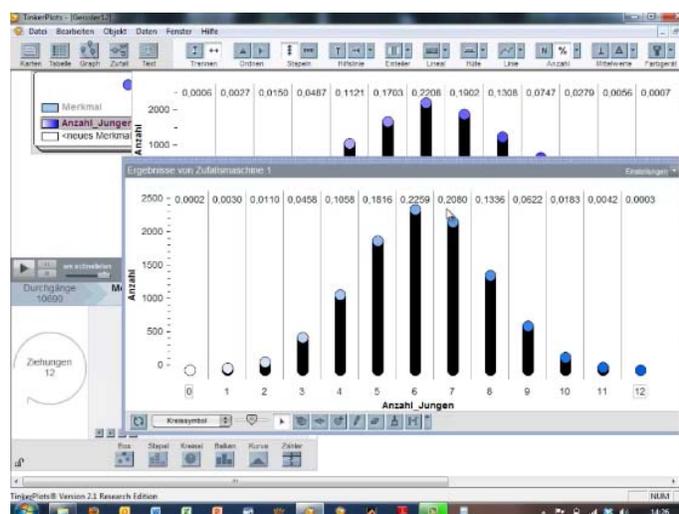
remembered the law of large numbers (it was in this context that the $1/\sqrt{n}$ law was taught). The pair wanted to use this law for a judgement. They rounded the number of boys to 10,000 and identified the interval for the simulated results as $\pm 1\%$ of the empirical data. So they viewed the frequencies in the empirical data as an estimation for having n boys out of 12 children. For bin 5 they noticed deviations between 1% and 3% and this did not fit with the $1/\sqrt{n}$ law. This was an appropriate use of the $1/\sqrt{n}$ law in this situation. After the application of the $1/\sqrt{n}$ law, they started to doubt their model. In the next transcript excerpt (see below) they set up a new model.

- 99 F I mean, is it right, from the biological point of view, that it is 50:50? So that exactly the half are boys and the other half are girls?
- 100 M I think, this is not considered here.
- 101 I These are two questions. We are still on task 3: you are comparing the distributions, and you use the $1/\sqrt{n}$ law and recognize that there are deviations between your simulated distribution and Geissler distribution. These are larger than we would expect via the $1/\sqrt{n}$ law. Have we summarized your thoughts correctly?
- 102 F Yes.
- 103 I So one can say, you are not totally satisfied with the model. Have I got you right?
- 104 M Yes.
- 105 I (Break, 5 sec) So we have to pose the question: is it really 50:50, for a boy or a girl birth? So you could change the probability for a boy or girl birth.
- 106 F Yes, we can!
- 107 M We could add more balls to the box of the Sampler, more balls labeled with boy and more balls labeled with girl. But we do not have enough information, we don't know anything about the preferences of boys and girls.
- 108 I You can do it. You have the Geissler distribution, an empirical distribution, and you know how many children have been born. 128000. You can read it above.
- 109 M Yes, 128280.
- 110 I So you also know how many boys have been born.
- 111 F Yeah! And then we can sum it up, so $7*12 + 60*11 + 298*10 + \dots$
- 112 I Yes, this is one possibility. Then you can see whether the probability of a boy birth is 0.5 or not.
- 113 M This is plenty of work to calculate the whole stuff!
- 114 *Francis and Marc start the calculation and calculate a boy's birth as summing up all boys born in Geissler data divided by 128280 as 51.54%.*
- 165 *On this basis they change their model (spinner in TinkerPlots Sampler) to $p(\text{boy})=0.5145$.*



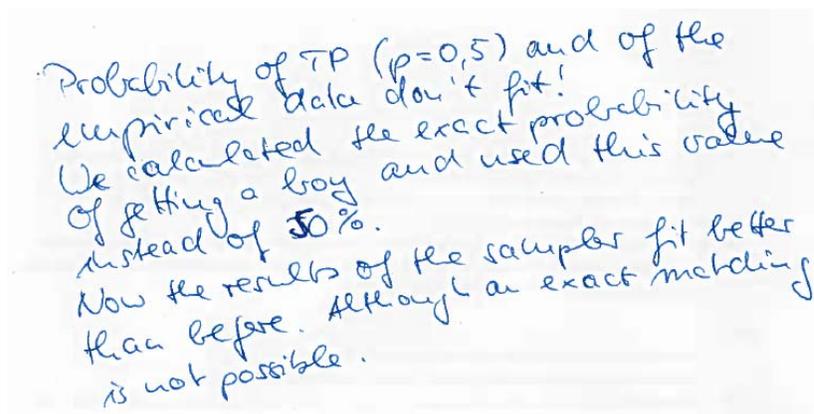
Line 99 is the first point at which Francis questioned the assumed probability for a boy's birth, but Marc did not go into this discussion. For the first time, context ("I mean, is it right, from the biological point of view, that it is 50:50? So that exactly the half are boys and the other half are girls?") played a role in their thinking. This was quite typical for most pairs, because none of them thought about the probability of a boy's birth for a very long time during their working processes. Marc's refusal to change the probability for a boy was the reason the interviewer intervened. In line 101 the interviewer summarized the students' findings and stated in line 103 directly, that the students were not satisfied with the model. Hereupon, Francis and Marc agreed. After a small break the interviewer directly intervened by proposing to change the probability for a boy and Francis agreed that they could change the assumed probability (line 106). Marc's first reaction was an undefined proposal to change the probability, but he did not have "enough information" to do this (line 107). After another intervention to prompt use of the Geissler data (line 108), Francis understood how to calculate the proportion of boys in the Geissler data and use it as an estimate for a boy's birth. At this point the interviewer was slightly wrong to call the boys' proportion in the Geissler data a probability. But this did not attract Marc and Francis' attention. The calculation took some time and after a while they ended with the calculated probability for a boy's birth as $p = 0.5154$ (with a small calculation error). They jumped into refining their model in TinkerPlots with the new probability. They took their model as it was and only changed the probability for boy to 51.45%. After plotting the new simulated distribution, they started comparing it to the Geissler data again. The plot was kept at the same place where it was before; only the frequencies in the plot changed due to the adjusted model. At this stage the next transcript excerpt starts.

- 166 M Okay. Then let's give it a try. Now we should be closer on the Geissler distribution. ...[They click on play on the Sampler and compare the frequencies of bin "5 boys"]
So, we are 1% here.



- 167 F Yes, a little bit more.
- 168 M Yes, a little bit more than 1% as well [Looking at bin “4 boys“].
- 169 F We can do it once again. Yes, it looks better at bin “5 boys”.
They click on play at the Sampler once again.
- 170 M We are more satisfied than before. All in all it fits better, because before we had larger deviations between Geissler data and our simulated data. We are more closer to our expected values. That’s good! Let’s continue!
- 171 F Oh, I did not see it [Geissler task].
- 172 M So, if we take this... It is possible to answer the main question.
- 173 F Do we have to read out the task?
- 174 I It’s enough to answer the leading question.
- 175 M So if we take this, ... [Answering the leading question...] then: YES.
- 176 F But not exactly. We cannot make it exactly for the frequency in each bin.
- 177 M No, but approximately. If we change the probability for boy or girl... if we know the probability for getting a boy or a girl in 12 children families, then we can model it with TinkerPlots approximately.
- 178 I Could you write that down, please?

Marc at first formulated his expectation “Now we should be closer on the Geissler distribution” (line 166). Then he looked at the relative frequency of bin 5, knowing from the comparison before that this was the bin with the largest deviation. He recognized a deviation of a little more than 1%. With a look at the frequencies of bin 4 and bin 5 Francis and Marc repeated the simulation. This was an adequate method to get again a feeling for the fluctuation of the frequencies. But they looked only at the two frequencies of bin 4 and bin 5 and only these frequencies were discussed. The small deviations led to their satisfaction with the new model. Here again (line 170), Marc formulated the frequencies in the Geissler data as their “expectation.” With no more comparing aspects, Marc summarized their conclusion about the overarching question in line 177, that they could model the process “approximately.” Their written answer (shown in Figure 14) showed that this was an important finding for them, that “an exact matching is not possible.”



Probability of TP ($p=0,5$) and of the empirical data don't fit!
 We calculated the exact probability of getting a boy and used this value instead of 50%.
 Now the results of the sampler fit better than before. Although an exact matching is not possible.

Figure 14. Written answer (translation by the authors) of Francis and Marc on the leading question

In sum, Francis and Marc showed a good understanding and could work through the task with only minor problems. They first set up their model with $p = 0.5$, applied the $1/\sqrt{n}$ law to compare the simulated data with Geissler data, and came to the conclusion that the model did not fit well to the data. Then they refined their model taking into account the given Geissler data and finally used the $1/\sqrt{n}$ law again to compare the simulated data with the Geissler data. The interviewer only gave five interventions (cf. Table 11) and the pair used the $1/\sqrt{n}$ law for comparing the distributions.

6.4. SUMMARY OF RESULTS

To conclude for this section, we summarize our findings with regard to the research questions.

“To what extent are the preservice teachers able to solve the Geissler task successfully?” Overall, the participants were able to solve the Geissler task. At expected stages interventions were necessary, but there were no additional difficulties. But, all pairs together needed 71 interventions. Most interventions were needed because they were satisfied with their model $p = 0.5$ and did not want to refine it without intervention by the interviewer.

“Which structure and phases can be identified when preservice teachers work on the Geissler task with TinkerPlots?” In our analysis we could identify statistical phases (see Table 7) and TinkerPlots phases (see Table 8). The processes of participants working on the Geissler task can be categorized into two types: Type one could be seen as setting up a first model with $p = 0.5$, comparing with the Geissler data and rejecting it, setting up the second model with $p = 0.5168$, comparing and accepting it. Type two incorporates one more cycle of setting up a model, comparing, and rejecting it. In the reasoning about model fit, fluctuation played a big role and our participants mentioned it at several stages. We think this was the main reason they agreed on model fit. Our participants knew that they would not get the exact same distribution and they knew that they needed to handle the fluctuation. In the participants' thinking the task was to find a model reproducing Geissler data and not to decide whether TinkerPlots could be used to model the underlying process.

“In which circumstances do the preservice teachers need support?” Interventions were necessary at expected stages, but there were no additional difficulties, so the interventions were well-prepared. Our participants showed good procedural knowledge of working with TinkerPlots. Very few interventions were needed at TinkerPlots-phases. Help was needed only for adjusting the bin width to 1 for creating the plot of the simulated data to look like the given Geissler plot. In the statistical phases more interventions were necessary than in the comparison phase.

“How do the preservice teachers reason about their models in the face of real data?” All pairs tried to compare the distributions via global features, three of seven pairs used the $1/\sqrt{n}$ law for comparing the Geissler distribution with a simulated distribution, and three of seven pairs used comparison “via deviation” and documented the directions of deviations in a table. To gain deeper insight we focused on the communication and argumentation of Francis and Marc when reasoning about their models. Their process can be seen as a good example of applying the $1/\sqrt{n}$ law for comparing the Geissler distribution with a simulated distribution taking into account numerical-percentile strategies to get the insight that a revision of their model would be reasonable for having a better fit between model and data.

7. DISCUSSION AND CONCLUSION

Because, as mentioned in the literature review, there are no empirical studies about working with the Geissler data in Germany, one first important implication is that the task of modeling a “family factory” with TinkerPlots was successful. We hence may state that elementary preservice teachers were able to set up models (factories) for given data with TinkerPlots successfully. They were able to run their models and compare the produced, simulated data in different ways. As Konold et al. (2007) postulated, a bridge between data and chance was built when working on our Geissler task.

Although data modeling was not an explicit issue in our MMDC curriculum, the participants were capable of using TinkerPlots for solving the Geissler task and doing data modeling. One small limitation is the comparison of distributions in TinkerPlots, because the heights of histograms do not necessarily correspond to the relative frequencies in each bin, which may hinder a graphical comparison because of the heights of the bars of the histograms. Apart from these problems, no procedural problems occurred in TinkerPlots. TinkerPlots was handled well by the participants: it helped them express, run, and change their models, and afterwards to collect and compare the data produced on the basis of this model. As stated by Pratt et al. (2011), TinkerPlots can help individuals reason about models. Participants saw TinkerPlots as valuable software to support the process of modeling. So we can say that Geissler data and the modeling of a family factory with TinkerPlots can bridge the concepts of data and chance and offer an application of data modeling with real data for statistics classrooms. Interventions can support learners with minimal contextual help so that they have the possibility of solving the task on their own to the greatest extent possible.

Regarding the empirical study reported in this article, one crucial finding was that all pairs started with an equal probability for a boy’s and a girl’s birth ($p = 0.5$) and that the participants did not think to question this seemingly common knowledge by reflecting more deeply about contextual issues in the beginning. Konold and Kazak (2008) reported that for this situation the “first model is a guess, an expectation or prediction” and that the model has to be refined later. Because adapting a model was not taught explicitly in the MMDC course, this was a new task to our students and therefore it was not surprising that their first

view of the simulated data led all pairs to the conclusion that the model with $p = 0.5$ fit the empirical data. In this case, the model has to be refined. Learners have to get a sense of which characteristics of models are crucial (e.g., independence or probability p in our case) and they have to get to know which components of a model can be improved. Furthermore, from a task design point of view, one might consider using a task with a situation having a more “natural” clash between data and model distributions that makes learners wonder and look for better models themselves (with needed interventions).

Apart from the problems adjusting the TinkerPlots plot with bin width = 1, problems arose most frequently in the comparison and evaluation phase (comparing the Geissler and the simulated distribution). A first comparison of Geissler data and simulated data was via shape, an approach that is also proposed by Konold and Kazak (2008) for doing such tasks with younger students. In our case we wanted participants to go beyond these strategies, and interventions were needed most often at this stage. The participants had problems with the interpretation of the deviations documented in the table and also had difficulties using the $1/\sqrt{n}$ law in this situation.

Another finding of this study is the structured process of the participants when working on the Geissler task. Like the well-known (and more general for mathematics education) modeling cycle, the participants went through the different steps of generating a model, running and evaluating it, and finally validating and (if necessary) revising it:

- Understanding the Geissler distribution
- Generating a model
- Setting up the model in TinkerPlots
- Comparing Geissler and simulated distribution
- Testing and refining first model

Further research may deal with teaching and design experiments about how data modeling and the application of open and real data, like Geissler data, can be implemented in secondary school statistics classrooms, and also used in the design of similar tasks.

REFERENCES

- Biehler, R. (2005). Authentic modeling in stochastics education: The case of the binomial distribution. In G. Kaiser & H.-W. Henn (Eds.), *Festschrift für Werner Blum* (pp. 19-30). Hildesheim: Franzbecker.
- Bliss, J., & Ogborn, J. (1989). Tools for exploratory learning. *Journal of Computer Assisted Learning*, 5, 37-50.
- Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. 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. 57-69). New York: Springer.
- Doerr, H., & Pratt, D. (2008). The learning of mathematics and mathematical modeling. In M. K. Heid & G. W. Blume (Eds.), *Research on technology in the teaching and learning of mathematics: Syntheses and perspectives. Mathematics learning, teaching and policy* (Vol. 1, pp. 259–285). Charlotte, NC: Information Age.
- Fisher, R. A. (1970). *Statistical methods for research workers* (14th ed.). New York: Hafner Press.
- Fisher, R. A. (1971). *The design of experiments* (8th ed.). New York: Hafner Press.
- Garfield, J. B., & Ben-Zvi, D. (2008). *Developing students’ statistical reasoning: Connecting research and teaching practice*. New York: Springer.

- Geissler, A. (1889). *Beiträge zur Frage des Geschlechtsverhältnisses der Geborenen. Zeitschrift des Königlich Sächsischen Bureaus*, 1-24.
- Griesel, H., Postel, H., Suhr, F., & Gundlach, A. (Eds.). (2003). *Leistungskurs Stochastik* [bearbeitet von H.K. Strick]. Hannover: Schroedel Verlag.
- Hadas, N., & Hershkowitz, R. (2002). Activity analyses at the service of task design. In A. D. Cockburn & E. Nardi (Eds.), *Proceedings of the 26th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 49-56). Norwich, UK.
- Harten, G. V., & Steinbring, H. (1984). *Stochastik in der Sekundarstufe I*. Köln: Aulis-Verlag Deubner.
- Kohlbacher, F. (2006). The use of qualitative content analysis in case study research. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, 7(1), Art. 21. Retrieved from <http://www.qualitative-research.net/index.php/fqs/article/view/75/153>
- Konold, C., Harradine, A., & Kazak, S. (2007). Understanding distributions by modeling them. *International Journal of Computers for Mathematical Learning*, 12(3), 217–230.
- Konold, C., & Kazak, S. (2008). Reconnecting data and chance. *Technology Innovations in Statistics Education*, 2(1). Retrieved from <http://www.repositories.cdlib.org/uclastat/cts/tise/>
- Konold, C., & Miller, C. (2011). *TinkerPlots 2.0*. Emeryville, CA: Key Curriculum Press. Available from www.tinkerplots.com
- Krummheuer, G., & Naujok, N. (1999). *Grundlagen und Beispiele Interpretativer Unterrichtsforschung*. Opladen: Leske+Budrich.
- Kuckartz, U. (2012). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*. Weinheim, Basel: Beltz Juventa.
- Leiss, D. (2007). *"Hilf mir es selbst zu tun" - Lehrerinterventionen beim mathematischen Modellieren*. Hildesheim: Franz Becker.
- Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, 8(1), 82-105. Retrieved from http://iase-web.org/documents/SERJ/SERJ8%281%29_Makar_Rubin.pdf
- Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahr, C. Knipping & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education* (pp. 365-380). New York: Springer.
- Pratt, D., & Ainley, J. (2008). Introducing the special issue of informal inferential reasoning. *Statistics Education Research Journal*, 7(2), 3-4. Retrieved from: http://iase-web.org/documents/SERJ/SERJ7%282%29_Pratt_Ainley.pdf
- Pratt, D., Davies, N., & Connor, D. (2011). The role of technology in teaching and learning statistics. 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. 97-107). New York: Springer.
- Riener, W. (2009). Soundcheck: CD contra MP3. Ein Hörtest als Einstieg in die Stochastik. *Mathematik lehren*, 153, 21-23.
- Samuels, M. W., & Witmer, J. A. (2003). *Statistics for the Life Science*. Upper Saddle River, NJ: Pearson.
- Voigt, J. (1984). *Interaktionsmuster und Routinen im Mathematikunterricht: theoretische Grundlagen und mikroethnographische Falluntersuchungen*. Weinheim: Beltz.
- Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3), 223-265.

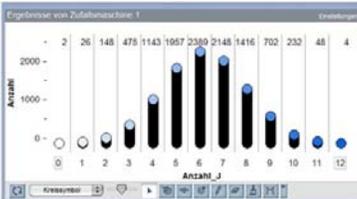
ROLF BIEHLER
University of Paderborn
Institute of Mathematics
Warburger Straße 100
33098 Paderborn
Germany

Appendix A: Statistical phases

	Statistical Phase	Explanation (What can occur in this phase?)	Example
D1	Describing and interpreting the given Geissler data	In this phase the participants describe or interpret the distribution of Geissler data via shape, symmetry, etc.	D: “<<Reading>> Explain “1821“ above bar five. I would say easily: These are 1821 cases, where five boys appeared amongst all 10690 families.” P: “Shape of distribution. I will call this symmetric. Approximately.” D: “Yes. And at six, at the half, there is the highest. The highest occurrence of boys.” (Denise-Paula, lines 6-8)
G1	Generating a model to reproduce data	In this phase the process of setting up the model is coded, e.g., talk about the model (probability for a boy’s birth, parameter of the model like number of children within each family, etc.) If there is talk about setting up the model in TinkerPlots, it also belongs to this phase.	R: “Okay. Now we should do a Sampler, yes? With male, female.” (Rosi-Ulla, line 15)
A1	Analyzing the results in TinkerPlots	In this phase the participants talk about technical aspects of producing and plotting data in TinkerPlots.	“Okay, and then count [number of boys] in join” (Sean-Susan, line 42)
C1	Comparing the sampled distribution with the Geissler distribution	In this phase the Geissler and the simulated distributions are compared - in the sense of one of the options a)-d). a) Looking at global features (symmetry, shape) and making an intuitive judgment whether the distributions are “similar” b) Document (write down) the frequencies in each bin, looking at deviations (difference	S1: “Perhaps, the deviations are smaller. But I don’t know.” (Sean-Susan, line 194)

		<p>between the bin frequency of simulated data and of Geissler data)</p> <p>c) Using the $1/\sqrt{n}$ law, respectively the \sqrt{n} law, to evaluate the size of the deviations in b. Using “goodness-of-fit” test.</p>	
V1	Validating the model	In this phase the participants validate the model based on the data generated by the model and the Geissler distribution. This leads to acceptance or revision of the model.	“Well, I would say, if we have a critical look at it, it doesn’t fit. The distributions seem to fit but the numbers [counts in each bins of the Geissler and simulated distribution] are different.” (Bert & Simone, line 163)
G2	Generating a revised model to reproduce data	In this phase the first model is revised. This includes setting up a revised TinkerPlots model as well.	“Yes, unless you change the probability for boys and girls. If we only knew the probability for getting a boy or a girl! So if we could, I don’t know, prove that it’s for example more likely to get a girl, then we could change the probability for getting a boy.” (Bert & Simone, line 164)
A2	Analyzing results in TinkerPlots again	Similar to phase A1.	No example
C2	Comparing the newly sampled distribution with the Geissler distribution II	Similar to phase C1 with new data of produced by the simulation of the revised model.	“Yes, it looks more alike now in bin 5 [counts of bin five in Geissler and new simulated distribution]” (Felix & Max, line 167)
V2	Validate the revised model	In this phase the participants validate the revised model based on data generated by the model and the Geissler distribution.	“We are more satisfied with our results now. Altogether it fits better than before, because we had larger deviations between the empirical distribution and this we built with TinkerPlots [Sampler]. We stay closely to the expected values.” (Francis-Marc, line 170)

Appendix B: TinkerPlots phases

	TinkerPlots Phase	Explanation (What can occur in this phase)	Example																																
S	Setting up the Sampler	<p>Every activity in the TinkerPlots Sampler belongs to this phase. e.g.</p> <ul style="list-style-type: none"> • set draw = 12 for 12 children within each family • set repeat = 10690 for the population of families • choose device that represents the assumption of the probability of getting a boy • Set speed of Sampler. 																																	
I	Identifying the random variable: Defining the result attribute “number of boys”	<p>The random variable used here is “number of boys”. This is a predefined attribute that operates on the “join”-attribute.</p> <p>In this phase the random variable is identified.</p>	 <table border="1"> <thead> <tr> <th>Formel</th> <th>Gesamt</th> <th>Anzahl_J</th> <th>Merkmal</th> </tr> </thead> <tbody> <tr> <td>10684</td> <td>M.M.J.M.</td> <td>5</td> <td>M</td> </tr> <tr> <td>10685</td> <td>M.J.M.J.J.</td> <td>8</td> <td>M</td> </tr> <tr> <td>10686</td> <td>J.J.J.J.J.</td> <td>10</td> <td>J</td> </tr> <tr> <td>10687</td> <td>J.J.J.J.J.</td> <td>9</td> <td>J</td> </tr> <tr> <td>10688</td> <td>J.J.M.J.M.</td> <td>9</td> <td>J</td> </tr> <tr> <td>10689</td> <td>J.J.M.J.M.</td> <td>6</td> <td>J</td> </tr> <tr> <td>10690</td> <td>J.J.M.J.M.</td> <td>7</td> <td>J</td> </tr> </tbody> </table>	Formel	Gesamt	Anzahl_J	Merkmal	10684	M.M.J.M.	5	M	10685	M.J.M.J.J.	8	M	10686	J.J.J.J.J.	10	J	10687	J.J.J.J.J.	9	J	10688	J.J.M.J.M.	9	J	10689	J.J.M.J.M.	6	J	10690	J.J.M.J.M.	7	J
Formel	Gesamt	Anzahl_J	Merkmal																																
10684	M.M.J.M.	5	M																																
10685	M.J.M.J.J.	8	M																																
10686	J.J.J.J.J.	10	J																																
10687	J.J.J.J.J.	9	J																																
10688	J.J.M.J.M.	9	J																																
10689	J.J.M.J.M.	6	J																																
10690	J.J.M.J.M.	7	J																																
P	Plotting the result attribute	All activities leading to the plot of the interesting attribute and changing the display belong to this phase.																																	
R	TinkerPlots Repeat	Running the Sampler in TinkerPlots once again.	See Paula & Denise, line 217																																
E	TinkerPlots Exploration	Non goal-oriented use (exploration) of TinkerPlots, which can not be covered by TinkerPlots Phases 1-3.	See Paula & Denise, line 176																																

Appendix C: Interventions

Main question: Does the TinkerPlots Sampler offer the possibility to model the distribution of gender in 10690 families with a random process, so that the distribution produced by this random process is nearly identical to the distribution of boys in the empirical data?

Interventions for Part 1: Understanding the distribution

What is the meaning of “1821” in the bin “5 boys”? Describe the shape of the distribution. Do you have assumptions about the shape of the distribution? Are you surprised about the distribution?

1_1 Intervention: Have a look at the distribution in detail. What can you say? Consider which types of shape you know from the course.

1_2 Intervention: Are you surprised at most families having 6 boys?

Interventions for Part 2: Setting up a family factory in TinkerPlots

Try to set up a model in TinkerPlots, which reproduces the distribution of gender of the families with twelve children in Saxony. Is it possible to produce a distribution via simulation that is nearly identical to the distribution in the Figure on the exercise sheet?

2_1 Participants do not find a starting point

2_a_1 Intervention: Maybe the simulation scheme will help you.

2_a_2 Intervention: In which way can you model the situation in TinkerPlots?

2_a_3 Intervention: What is the probability for a boy’s birth?

2_a_4: Intervention: *Help with the simulation directly.*

Interventions for Difficulties with TinkerPlots

2_b_1 Difficulties with “draw”

2_b_1_1 Intervention: Have a look at the Sampler in detail.

2_b_1_2 Intervention: Have a look at the number of draws.

2_b_1_3 Intervention: Due to the number of 12 children within each family, you need 12 draws.

2_b_2 Difficulties with “repeat”

2_b_2_1 Intervention: Have a look at the Sampler in detail.

2_b_2_2 Intervention: Have a look at the number of repeats.

2_b_2_3 Intervention: Due to the number of 10690 families, you need 10690 repeats.

2_b_3 Difficulties with the result attribute (contextual)

2_b_3_1 Intervention: Think about what you want to display.

2_b_3_2 Intervention: Compare with the empirical distribution. Which attribute is displayed there?

2_b_3_3 Intervention: You have to display the number of boys per family.

2_b_4 Difficulties with the result attribute (technical)

2_b_4_1 Intervention: *Help with the technical aspects directly.*

2_b_5 Difficulties with the plot

2_b_5_1 Intervention: Produce a distribution that is nearly identical.

2_b_5_2 Intervention: *Help with the technical aspects directly.*

Interventions for Part 3: Comparing distributions

Compare the distributions of the simulated and the empirical distribution. What do you see?

3_1 Nothing attracts attention: Participants think that their model fits

3_1_1 Intervention: Why do you think so?

3_1_2 Intervention: Repeat the simulation a few times.

3_1_3 Intervention: Have a look at single bins, for example at bin 5 and observe the deviations.

3_1_4 Intervention: What about the deviations generally?

3_1_5 Intervention: Do the deviations look random? Use the table.

3_2 Attract attention to: Participants observe deviations: Repeat is not 10690

3_2_1 Intervention: Have a look at your Sampler.

3_2_2 Intervention: How many families do you want to produce?

3_2_3 Intervention: You need 10690 repeats.

3_3 Attract attention to: Participants observe deviations that are “too large” (nonspecific)

3_3_1 Intervention: Explain this a little more.

3_3_2 Intervention: Try to use percentages instead of absolute frequencies.

3_3_3 Intervention: Try to remember how we proceed in our course in the case of accuracy of simulation.

3_3_4 Intervention: Maybe the $1/\sqrt{n}$ law can help you.

3_4 Attract attention to: Participants observe deviations that are “too large” but do not recognize a pattern

3_4_1 Intervention: Compare bins with too many or too few boys in contrast to the empirical distribution. Use the table.

3_4_2 Intervention: Repeat the simulation a few times.

3_4_3 Intervention: What about the deviations?

3_4_4 Intervention: Do the deviations look random? Use the table.

Intervention_Z (Recall): Are you satisfied with your model?

Interventions for Part 4: In case you are not satisfied with your comparison, adjust your model and simulate again.

How can you change your assumptions for part 2 in a way that the data produced by the model do fit better to the empirical data?

4_1 Intervention: Which assumptions can you modify?

4_2 Intervention: What probability for a boy’s birth did you assume?

4_3 Intervention: What is the relative frequency for a boy’s birth in the empirical data?