

On middle-school students' comprehension of randomness and chance variability in data¹

Joachim Engel, Hannover (Germany)
Peter Sedlmeier, Chemnitz (Germany)

Abstract: Understanding variability in empirical data is at the core of statistical reasoning and thinking. Of particular interest is how students' comprehension of chance and variability develops over time. This article reports the results of a cross-sectional study that examined how students' statistical literacy evolves with increasing age. Our results are discussed and related to earlier studies with children by Fischbein and Green and with adults by Sedlmeier. Our study replicates in a modified form earlier investigations in other countries and confirms for German students conclusions from earlier studies. In particular, there are no indications of an improvement with increasing age. Our findings are consistent with findings in judgment research.

Kurzreferat: Im Zentrum statistischen Denkens steht der Umgang mit Variabilität in empirischen Daten und die Fähigkeit, mit Hilfe des Zufallsbegriffs nicht erklärte Variation in Daten zu modellieren. Die vorliegende Studie berichtet von einer Querschnittsuntersuchung von 222 Schülern der Klassenstufen 5, 7 und 9 zur Entwicklung des Verständnisses von Zufall und Variabilität. Dazu wurde anhand geeigneter Aufgaben untersucht, wie stark statistische Kompetenz bei Schülern ausgeprägt ist und ob sie sich mit zunehmenden Alter der Schüler verändert. Die Ergebnisse werden zu vergleichbaren früheren Untersuchungen bei Kindern von Fischbein und Green und bei Erwachsenen von Sedlmeier in Beziehung gesetzt. Die Studie repliziert in modifizierter Form frühere Untersuchungen in anderen Ländern und bestätigt für deutsche Schüler woanders gefundene Resultate. Insbesondere liegen keinerlei Anzeichen für eine Verbesserung des Verständnisses von Zufall und Variabilität mit zunehmendem Alter der Schüler vor. Die Ergebnisse sind konsistent mit Befunden aus der Urteilsforschung bei Erwachsenen.

ZDM-Classification: K50, C30

1. Chance, Randomness, Statistical Reasoning and Intuition

"In life nothing is certain –except death and taxes"

This quotation by Benjamin Franklin (1706-1790, US president) indicates that uncertainty is a condition of human existence. In our daily lives we are continually confronted with situations where we have to respond, (re-)act and make decisions without having complete information and insight available to predict with certainty the consequences of whatever we do. In cultures and eras in which the concept of probability had been poorly developed, people tended to resort to

inadequate explanations and inappropriate attempts to give meaning to whatever they experienced by attributing it to superstition, magic or moral meanings (punishment or rewards for committed actions). Such attributions made sense of whatever happened to them, when no immediate and compelling causal reasons could be identified. In the following we report an empirical study conducted in the fall of 2002 that investigates how uncertainty is dealt with in a highly educated society. In particular, we investigated which statistical competencies are available to school students and how these competencies evolve over the years.

Some of the results from developmental psychology create optimistic expectations. Various studies indicate that children have valid intuitions about probability. Already in elementary school, children are sensitive to the effects of randomness (Kuzmak & Gelman, 1986). Around the age of 8 most children already know that conjunctive sets (e.g. flowers which are yellow and primroses) cannot be larger than components (in the example: primroses) (Inhelder & Piaget, 1959/ 1964). Already at age 11 children show some understanding for the empirical law of large numbers (Piaget & Inhelder, 1951/ 1975). All these results indicate an increasing and progressive improvement in statistical judgment. In the literature on the statistical judgment of adults, however, the picture changes dramatically: numerous examples demonstrate that adults have severe difficulties when dealing appropriately with aspects of probability, such as the conjunction of probabilities or the role of the sample size in the empirical law of large numbers (Kahneman, Slovic & Tversky, 1982; Piattelli-Palmarini, 1994). In the meantime this pessimistic perspective has been somewhat revised (e.g., Gigerenzer, 1996), but for some kinds of problems erroneous judgments have been replicated over and over again. Why do adults have these problems? Could it be that school education is responsible for the deterioration in statistical reasoning? Indeed, there are some indicators for this provocative assertion. When students start attending school they have valid primary intuitions regarding randomness, but school appears to be detrimental to these intuitions because of the deterministic world-view advocated by many teachers:

"The child is taught [in school] that explanation consists in specifying a cause; that a scientific prediction must be a certainty; that ambiguity and uncertainty are not acceptable in scientific reasoning and so on. Even if all this is not explicitly stated, it is implied in all that is taught in school."
(Ephraim Fischbein, 1975, p. 71)

Fischbein based his conclusion on studies about the statistical intuitions of school students aged between 5 and 14. Further evidence for misconceptions in students' statistical judgments is reported by Shaughnessy (1992), Batanero et al. (1994) and Green (1982, 1986, 1990).

Does school education nourish appropriate intuition about randomness or do our schools rather support inappropriate deterministic causal thinking as

¹ Revised version of a German version of the paper by the same authors in: *Unterrichtswissenschaft. Zeitschrift für Lernforschung*, 2, 2004, pages 169-191

Fischbein’s quote suggests? If the latter is the case, then the performance of our students on certain tasks should not improve with the years. By focusing on middle school students (aged between 10 and 16) we made our participants’ age range comparable to that in earlier studies (e.g., Green, 1982)

2. Task selection

In selecting our tasks caution was called for to avoid confounding general mathematical abilities that increase with the years of attending school and special competencies in dealing with random phenomena which were at the center of our study. Therefore, we excluded tasks whose solution rates correlate with increasing general mathematical skills (like dealing with proportions or fractions). Furthermore, to compare our study with previous studies and generalize from them, we chose tasks or variants of tasks that had been used in previous investigations. The selected tasks deal with the generation of a typical random sequence, the concept of a planar random distribution, the empirical law of large numbers and the probability of conjunctive events.

2.1 Random sequences

The concept of independence of events is an idea many people have difficulties grasping. Even if independence is very easy to define mathematically by the property that a joint probability of two events is the product of the two single event probabilities, research in cognitive psychology (e.g., Nickerson 2002) points to massive problems in human understanding. Steinbring (1986) suggests that the concept of independence is of a theoretical nature, whose application in concrete situations creates problems. A famous example is the “gambler’s fallacy”. When tossing coins, many people believe that after various preceding events of “heads” the probability of “tails” is increased. In order to survey students’ understanding of independence of a random sequence, we asked them to make a judgment about a typical sequence of coin tosses.

Task 1: Imagine you are flipping a coin twenty times. Each throw has either “heads” (H) or “tails” (T) showing on top. What could the result of 20 flips look like? Write down a sequence of “heads” and “tails” that looks typical for you.

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

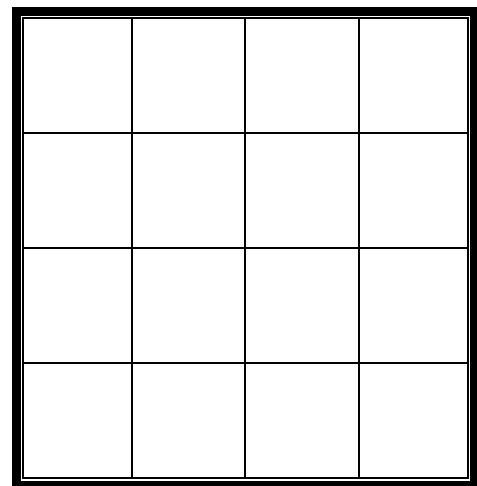
2.2 Random distributions

While the last task deals with the temporal sequence of events, the next task looks at the planar distribution of random events. Despite global regularity, random events are individually unpredictable. How do students decide on the cognitive conflict between a global uniform distribution and the unpredictable behavior of single elements in a small sample?

The following task stems from Piaget and Inhelder (1975) and has been used with school students in different variants by Green (1982, 1986, 1990). The starting point is an everyday phenomenon that already children observe: when it starts snowing, how do the snowflakes distribute within a small area of manageable size? The cognitive conflict is provoked by the unforeseeable behavior of single flakes on the one hand and the well-experienced fact that on the other hand after some time of snowfall the snow is approximately of even height, for example, on a flat rooftop. The surprising findings of Piaget and Inhelder, and Green are not only that most students chose a strictly regular distribution that has nothing to do with randomness, but in particular that the tendency to strict regularity increases with students’ age. Green (1982) concludes that there are two opposing tendencies at work here: maturation and increasing experience versus the dominance of mechanistic-deterministic thinking that attempts to give strictly causal explanations for everything and thus allows no space for taking account of randomness.

In contrast to Green, whose tasks were in a multiple choice format with several patterns to choose from, we represented this task in an open format, in order to allow for a variety of possible answers. Our objective was to find out how children decide between random variation and a global uniform distribution of the flakes. We also requested a verbal explanation to gain additional insights into the students’ cognitive world.

Task 2: The square rooftop of garden shed has 16 equal-sized square-tiles.



It starts snowing. After a short while 16 snowflakes have landed on the roof. Please, note down where the 16 flakes could have landed (mark an x for a flake). Explain your answer:

2.3 Empirical Law of Large Numbers

In the literature on how sensitive children and adults regard the influence of the sample size in statistical judgments, one encounters a remarkable inconsistency: According to Piaget and Inhelder (1975) children at age 11 or 12 already show a clear sensitivity for the

influence of sample size while until recently the dominant opinion in judgment research was that adults ignore sample size in relevant judgments (e.g. Kahneman & Tversky, 1972; Piattelli-Palmarini, 1994). A closer look, however, reveals that in judgment research two different types of tasks have been used (Sedlmeier, 1998; Sedlmeier & Gigerenzer, 1997): on the one hand, the tasks were *frequency distributions* tasks, where confidence judgments about proportions or means from a sample (a frequency distribution) were asked for. On the other hand *sampling distribution* tasks were used, where the variation of empirical sampling distributions has to be judged. Frequency distribution tasks (to which the tasks used by Piaget and Inhelder are to be subsumed) were generally solved correctly. This is not, however, the case for sampling distribution tasks. Sedlmeier and Gigerenzer (1997) explain this by pointing out that humans have an intuition at their disposal that corresponds to the empirical law of large numbers, which is applicable to frequency distributions but not to sampling distributions (see Sedlmeier, 1999, 2002, for an explanation based on learning theories). The objective of the present study was to investigate whether the differences in solution rates between frequency distribution and sampling distribution tasks persist even among children and juveniles and whether the sensitivity to the influence of sample size changes within the course of the school years. The tasks used were variants of the “Maternity-Ward” problem of Kahneman and Tversky (1972). The sampling distribution version of the task was:

Task 3: *About half of all newborns are girls, the other half boys.*

*Hospital A: an average of 3 children are born every day
Hospital B: an average of 5 children are born every day
Which hospital will have more days within the course of a year where all the births are baby girls?*

- Hospital A (with 3 births a day)*
- Hospital B (with 5 births a day)*
- In both hospitals chances for this to happen are about equal*

In the frequency distribution version students received the following question instead:

In which hospital is it more likely that on a certain day all births are baby girls?

The frequency distribution in this task relates to the distribution of boys and girls in a sample, i.e. the births on a day while the sampling distribution consists of the 365 portions of baby girls throughout a year.

2.4 Probability of Conjunctive Events

The literature reports quite diverging results relating to judgments on probabilities of conjunctive events. Inhelder and Piaget (1964) found, as mentioned above, that children by the age of eight are capable of recognizing that conjunctive sets cannot be larger than the components. Once again, initiated by the results of

an often replicated study by Tversky and Kahneman (1983), the dominant opinion is that adults are generally not capable of estimating the probability of conjunctive events properly. Hertwig and Gigerenzer (1999), however, were able to show that erroneous estimates of probabilities of conjunctions were caused by semantic ambiguities (e.g. how the word “and” was understood). A presentation of the task information in a frequency format where the ambiguity had been avoided (as in Inhelder & Piaget, 1964) led to solution rates of up to 100%. However, there is no consensus about the concrete mechanisms of how the frequency formulation works (Mellers, Hertwig, & Kahneman, 2001). In our study we used a task with the same structure as in several previous studies in judgment research. In addition to the format (probability versus frequency) we varied the temporal perspective (see below). In the (usual) probability version the task (version A) reads as

Task 4: *A student earned a D in math on his report card at the end of the school year. Which of the following statements are more likely for him to be true?*

- He had an F in his mid-school-year report card.*
- He had received tutoring in the second half of the school year and earned an F in his mid-school-year report card.*

In the corresponding frequency version the tasks were formulated as follows (Version B)

30 students in a class earned a D on their end-of-the-school-year report card. Which of the following statements describe more of these students?

- They earned an F in their mid-school-year report card.*
- They received tutoring in the second half of the school year and earned an F in their mid-school-year report card.*

The reason for varying the temporal perspective was the observation that semantic ambiguities may be bigger when the statement to be judged is *before* the description of the problem (as in Version A and B), than when these statements refer to future (from the time of problem description) events. Accordingly, the “future oriented” versions were in version C:

A student had an F in math on his mid-school-year report card. Which of the following statements are more likely for him to be true?

- He had a D in his report card at the end of the school year.*
- He received tutoring in the second half of the school year and had a D in his report card at the end of the school year.*

In version D:

30 students in a class earned an F on their mid-school-year report card. Which of the following statements describe more of these students?

- They had a D on their report card at the end of the school year.*
- They received tutoring and had a D on their report card at the end of the school year.*

3. Method

3.1. Participants

To test our hypotheses we gave the test tasks to 222 school students attending grades 5, 7 and 9 in the three main branches of the German school system in September of 2002 in the Stuttgart area. The German school system (while varying from state to state) separates students from grade 5 (7) on according to academic achievements into Hauptschule, Realschule and Gymnasium. We presented the four tasks discussed in Section 2 to an entire class at the above grade levels and school types. The exact composition of the sample is given in Table 1. In respect to age and sex the classes of the same grade level did not differ essentially. The average age was 10.6 years for 5th graders, 12.7 for the 7th graders and 14.9 for the 9th graders. Altogether, we tested 106 boys and 116 girls. In respect to scholastic achievement (relating to mathematics) we selected classes that were – according to their teachers’ judgment – regarded as average, relative to their corresponding school type. The three categories of schools were chosen to have a sample that can be considered representative for all students. In the German state of Baden-Württemberg students attend these three types of school in roughly equal percentages. The students had been instructed in mathematics according to the state curriculum. As stochastics in Baden-Württemberg (in 2002) has been taught only at the Gymnasium (beginning in grade 10), it is safe to say that the students had no specific classroom instruction and no previous experiences with random experiments and simulation.

Table 1. Composition of the sample

	Haupt- schule	Real- schule	Gymna- sium	Sum
Grade 5	25	24	27	76
Grade 7	23	26	28	77
Grade 9	25	25	19	69
Sum	73	75	74	222

Unknown to the students, the different versions of tasks 3 and 4 were randomized among the participants.

4. Results and Discussion

Our focus was on the change in statistical literacy over the years of attending middle school. The main results in each task refer to the trend from grade 5 to grade 9. We didn’t differentiate between various school types, because in the Hauptschule as well as in the Gymnasium selection effects occur: strong Hauptschule students and poor Gymnasium students move to other school types over the years. In the Hauptschule, with increasing grade level we find a large proportion of students who do not meet the criteria necessary for a transfer to an academically more demanding school, while selection at the Gymnasium (and partially as well at the Realschule) will work in the opposite direction: Students not meeting the expectations of the school are

transferred into the Realschule or Hauptschule. Therefore, an isolated consideration of the course of students within one type of school only would be difficult to interpret. Nevertheless, we analyzed the data with respect to systematic differences regarding school types and report results in those cases where we observed differences.

For analyzing the first two tasks - the production of a random sequence and random distribution – several methods are available (e.g., Nickerson, 2002; Green, 1990). Therefore, we give the analysis of these tasks more space than the tasks referring to the law of large numbers and the probability of conjunctive events.

4.1. Production of random sequences

Students’ responses to task 1 were analyzed according to four different criteria: the relative frequency of “heads”, the relative transitional frequency, the number of runs and the length of the longest run.

a. Distribution of the number of “heads” within a coin flipping sequence of length 20

A comparison of the mean values for the number of “heads” for all grade levels has been done to check if students have a preference for either one of the two possible results or if they decide in about equal numbers for “heads” and “tails”.

Table 2. Average selected number of “heads”. The values in brackets are averages excluding two outliers.

	Grade 5	Grade 7	Grade 9	Total
Hauptschule	9.96	9.87	10.04	9.96
Realschule	9.75	10.8	10.2	10.01
Gymnasium	9.26 (9.62)	10.0 (10.26)	9.73	9.66 (9.89)
Total	9.64 (9.77)	9.98 (10.08)	10.01	9.88 (9.95)

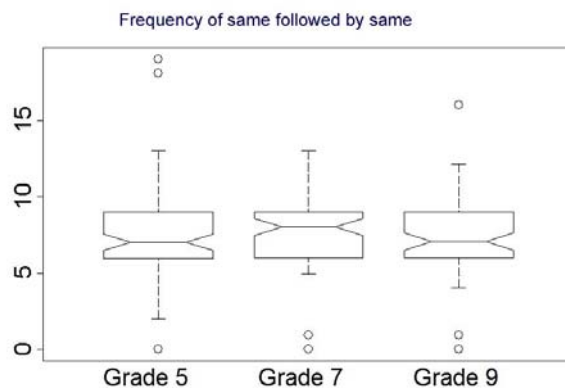
We conclude from Table 2 that students throughout all grade levels and school types produced “heads” and “tails” in about equal numbers, and are close to the theoretically correct value of 10. If we drop one fifth-grader from the Gymnasium, whose sequence contained 20 “tails” and a seventh-grader (also Gymnasium) with a sequence of 17 “tails” from the sample, the average values are even closer to the value of 10. As median we obtained 10 throughout all grade levels and schools. Hence, there is no significant preference for “heads” over “tails” or vice versa. Green (1990) computes in his similar investigations the standard deviation and concludes that the standard deviations are consistently below their theoretically correct value of $\sqrt{20 \cdot \frac{1}{2} \cdot \frac{1}{2}} \approx 2,23$. From this he concludes that the random variability is underestimated, i.e. students choose 9, 10 or 11 “heads” too often, i.e. numbers which are close to the expected value of 10. We find this conclusion questionable, because each of the approximately 25 students in a class has been asked to produce a *single* typical sequence – a task conceptually different from a

request to a single student to produce 25 typical random sequences. While the latter design may be appropriate to measure the individually perceived variation, we find it rather logical that, in response to our (and Green's) task, most students produce a coin flipping sequence with the maximum likely number of "heads".

b. *Analysis of relative transition frequencies: How often is one side of the coin followed by the same side?*

Whereas the last criterion referred to the global number of "heads", we focus now on the sequential occurrence of "heads" and "tails". An important consideration here is the independence of the single throws. As a coin has no "memory" the probability for "heads" is 0.5 at each throw, regardless of the history of previous outcomes. At this point human intuition fails for many people. After five times "heads" in a row, a majority of adults – as documented in numerous studies – expects the probability of "tails" to be up in the next throw ("Now it is the turn for tails"). Human intuition struggles obviously with the concept of independence, represented here by the irregularity of a random sequence. In order to evaluate the comprehension of the irregularity of a random sequence, we computed for each student the transition frequencies of two like symbols in consecutive order ("heads" following "heads" or "tails" following "tails"). For example, the eight symbol sequence HHTHHTTT contains seven transitions, among them four transitions from a symbol to the same symbol, namely twice H followed by H and finally two T transitions in the triple occurrence TTT at the end. Hence, we observe a frequency for "same followed by the same" of 4. Independence here is represented by a value expected to be somewhere close to 3.5, or a relative frequency close to 0.5. A coin flipping sequence of length 20 has 19 transitions. Under the independence assumption we compute an expected value of same-to-same transitions as $19 \cdot 0.5 = 9.5$. Figure 1 displays boxplots of the transitional frequencies separated by grade level. There are barely any differences to discern among the grade levels. A comparison of the average transitional frequencies leads to the same conclusion: while the coin flipping sequences of the 7th graders average out at 8.01, the corresponding values for 5th graders were 7.58 and or the 9th graders 7.48. What is the percentage of students here that is below the theoretical expected value? In grade 5 this amounts to 80%, in grade 7 to 83%, and even 84% of the 9th graders chose sequences which switch too often, i.e. the same being followed by the same occurs too frequently. Despite this slight decrease over the years, the observed differences are not significant in the sense of statistical inference.

Figure 1. Boxplots for the frequencies the same ("heads", "tails") ... followed by the same



c. *Number of Runs*

While all single realizations of a coin tossing sequence occur with the same probability

[here with $P = \left(\frac{1}{2}\right)^{20}$],

the randomness of a concrete sequence can be judged by a pattern analysis. A consecutive sequence of the same result is called a run, e.g. the sequel THTTTHHH consists of 4 runs (first a T Run of 2 T is followed by an H run, then again a T run of length 2 and finally an H run containing 3 Hs). In random sequences the occurrence of very many as well as very few runs is unlikely. Probabilistic considerations lead to the conclusion that the expected number of runs in a coin flipping sequence of length 20 is 10.5. Figure 2 shows boxplots separated for each grade level and Table 3 gives the empirical means of the number of runs.

Figure 2. Boxplots for the number of runs

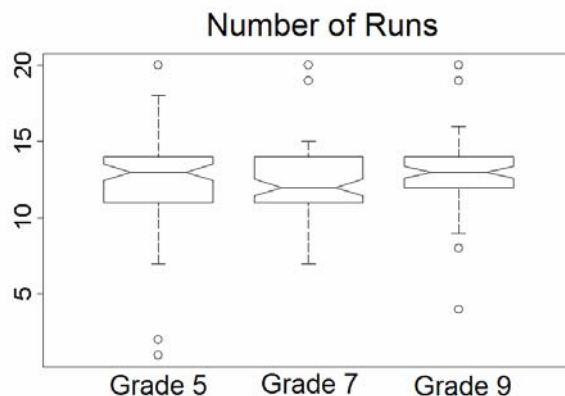


Table 3. Means for the number of runs

	Grade 5	Grade 7	Grade 9	Total
Hauptschule	13.44	12.13	12.96	12.86
Realschule	13.75	12.50	12.28	12.83
Gymnasium	10.78	12.14	12.42	11.72
Total	12.59	12.26	12.55	12.47

We obtain essentially the same picture as in the previous section. Regardless whether we take the average values or the more robust medians and percentiles (see Figure 2) we notice: while among the

school types students from Gymnasium slightly outperform the other two student populations, there are no statistically significant differences among the grade levels. The coin flipping sequences have too many runs, i.e. more than the theoretically expected number. Among the 5th graders about 80% of the students list too many runs, in grade 7 it is 82% and among 9th graders even about 85% of all students have too many runs.

d. *Length of the longest run*

Another widely used criterion to judge the randomness of coin tossing sequences is based on the length of the longest run. As a result of false understanding of independence many adults estimate the length of the longest run as too short. Theoretical considerations (see, for example, Eichelsbacher, 2002) lead to the conclusion that the expected length of the longest run is approximately given by the base 2 logarithm of the sequence length: $L_n \approx \frac{\log(n)}{\log(2)} = \text{ld}(n)$.

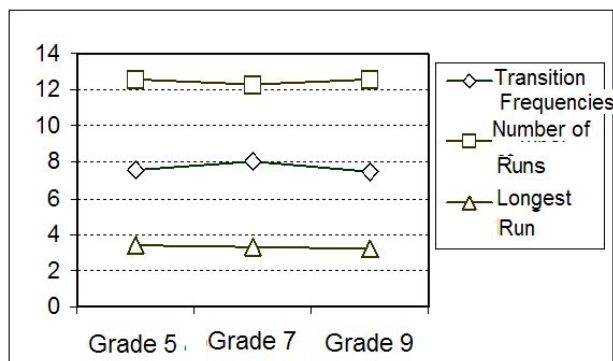
In the case of $n = 20$ we compute a value of $L_n \approx 4.322$, i.e. a run of length 4 or 5 is not at all unusual in a sequence of length 20, but quite within the range of what is expected of a random sequence of that length. Table 4 shows the empirical averages of the sequences produced by students. A boxplot here is not meaningful as *all* medians and *at the same time all* lower quartiles amount exactly to 3. Here, we also have an advantage for the students from the Gymnasium whose mean is closer to the theoretically expected value of 4.322, while students from the Realschule score even slightly worse than students from the Hauptschule. However, differences among the various grade levels are very small.

Table 4. Means of the length of the longest run

	Grade 5	Grade 7	Grade 9	Total
Hauptschule	2.80	3.39	3.24	3.14
Realschule	2.88	3.15	3.16	3.06
Gymnasium	4.41	3.36	3.21	3.70
Total	3.39	3.29	3.20	3.29

Summarizing, we conclude: in simulating coin tossing sequences, students in all kinds of schools and throughout all grade levels are very precise in balancing the number of “heads” and “tails”. Students’ problems start, however, when they are asked to represent independence adequately. This is also true for all grade levels. There are no significant changes (either up or down) over the years. The longest runs are not long enough. Difficulties with representing independence are also reflected in the transition frequencies, that is, in judgments of how often an outcome is followed by the same outcome. The transition frequencies are consistently smaller than 0.5. Students from the Gymnasium have fewer difficulties with these concepts than students attending the Realschule or the Hauptschule. The grade levels barely differ at all.

Figure 3. Display summarizing results for Task 1. The theoretically derived benchmark values are: transition frequencies = 9.5, number of runs = 10.5, longest run = 4.322



4.2 Comprehension of Planar Uniform Distribution

Piaget and Inhelder (1975) were the first to investigate children’s comprehension of patterns in random sequences. They designed a scenario, where raindrops were falling onto the slabs of a paved road. The desire for regularity dominated the children’s prediction. Asked where the next raindrops were about to land, 6 to 9 year old children distributed the drops in about equal numbers on each squared pavestone. When drops had been falling on all but one square pavestone, the children inevitably placed the next drop on that empty pavestone. With increasing age this type of thinking disappeared and irregularities in the distribution were more accepted. Piaget & Inhelder (1975) concluded that children were entering a phase where they started to get some comprehension of the law of large numbers that at the same time explains global regularity and local variation in the experiment. Green (1986) discussed this theory, whose study of 11 to 16 year old children indicated that the percentage of children decreased with age, who recognized random and “half-random” distributions. Green (1982) presented 2930 British students aged between 11 and 16 the raindrop task in multiple-choice format (four possible patterns to choose from were given), and discovered no sign of improvement with increasing age. In another paper (Green, 1986) he reports the results of the raindrop task with 1600 children of age 7 to 11 and concludes that there is no improvement when comparing these age groups, with the only exception being the youngest group of 7 to 8 years old who clearly perform worse. Also Borovcnik and Bentz (1991) discuss this task and question the appropriateness of the given answers for assessing the effects of chance.

We picked up this type of question in task 2, but in contrast to the task used by Green our students were free to choose their own distribution of snowflakes on the garden shed. In addition we asked them to explain their answer in their own words. When evaluating student responses we identified four different patterns:

1. “Strictly determinist”: Fixed, rigid patterns for how the flakes are placed are clearly discernible. For examples, each flake is located exactly at the center of a tile, and no tile is left out.
2. “Moderately determinist”: Patterns are without doubt recognizable, even though the distribution of

flakes is not determined down to the last detail, e.g., each tile contains a flake that is placed somewhere on that tile, or is positioned on tiles that have a particular position such as all interior or all the outer tiles of the roof.

3. "Novice": It is obvious that the flakes are placed somewhere on the roof. Patterns are not perceivable. In particular, there is at least one empty tile. However, the variability in the number of flakes per tile is very small. There are some empty tiles and other tiles with more than one flake. A maximum of three tiles remain empty.
4. "Expert": Seemingly a random distribution. There are no recognizable patterns; between four and eight tiles remain empty. There is no symmetry or any other discernable pattern or regularity. Flakes are placed haphazardly somewhere on the rooftop.

The classification of the student responses into one of these four categories has been done by three experts working independently. In a few doubtful cases the students' written explanation was considered. Altogether there remained only one student who could not be classified. The quantitative evaluation has been done according to three different criteria (Score 1, Score 2 and Score 3), that were defined as follows:

Score 1 = relative number of those students, who were not strict determinists

Score 2 = relative number of experts or novices

Score 3 = relative number of experts

The requested verbal explanations and reasoning proved to be quite revealing as regards to students' difficulties in representing randomness appropriately. Here are some examples of the position of the determinists:

- *Because we have 16 tiles and 16 flakes, therefore [exactly one flake] falls onto each tile* (Hauptschule, Grade 5)
- *Because exactly one flake falls onto each tile* (Hauptschule, Grade 7)
- *The flakes fall uniformly* (Realschule, Grade 7)
- *Because flakes fall everywhere in equal numbers* (Gymnasium, Grade 7)
- *Theoretically, the 16 flakes have to be distributed equally* (Realschule, Grade 9)
- *Because it's snowing uniformly* (Gymnasium, Grade 9).

The following episode was observed during the test. While being circumstantial, we think it is revealing of the exclusive prevalence of deterministic thinking in the classroom: In a seventh grade classroom the teacher identified to the investigator his two "best students". Their responses to the snowflake task were distinguished by their marking their crosses with particular accuracy in the center of the 16 tiles.

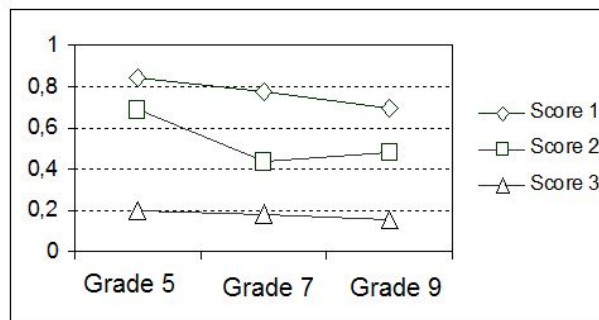
In responses that represented an appropriate comprehension of randomness, students recognized the unpredictability of chance events or expressed the influence of uncontrolled third variables like wind or temperature as a cause of irregularity:

- *Because the wind make the flakes fall on these spots* (Realschule, Grade 5)
- *Because in Science nothing ever is perfect. Therefore, I left 3 tiles empty.* (Gymnasium, Grade 5)
- *Because snowflakes fall once here, once there* (Hauptschule, Grade 7)
- *It depends on how the wind is blowing.* (Hauptschule, Grade 7)
- *Snowflakes fall without any system* (Gymnasium, Grade 7)
- *Snow doesn't just simply fall. It is not controllable or predictable.* (Hauptschule, Grade 9)
- *You never know where the flakes will land. And the chance that 16 flakes land on all 16 tiles is very low* (Hauptschule, Grade 9)
- *There is no reason, they fall randomly. I closed my eyes and marked 16 dots with my pen.* (Gymnasium, Grade 9)

Considerations of the students' quotations provide insight into the students' cognitive world, how they make sense of and reason about randomness. These quotations, however revealing they may be, are not suited to identify trends in students' thinking through their year of school attendance. This is done by looking at the three scores defined above. An analysis of the scores does not reveal any systematic difference between the three school types while deterioration over the years becomes obvious from visual inspection of the data (see Figure 4). Green (1982), by presenting in a multiple choice format patterns that roughly correspond to our categories of "strict determinist", "moderate determinist" and "expert", arrived at exactly the same conclusion. He commented on the result that the percentage of "expert" solutions dropped from 26 % for the 11 year old to 18 % for the 16 year old as follows:

"The astonishing thing about this task is that performance declines with age. ... It can be hypothesized that we see at work here two opposing tendencies – maturation / experience on the one hand and dominance of mathematical / scientific deductivism on the other which stifles the appreciation of randomness by seeking to codify and explain everything."
(Green, 1982, S. 744)

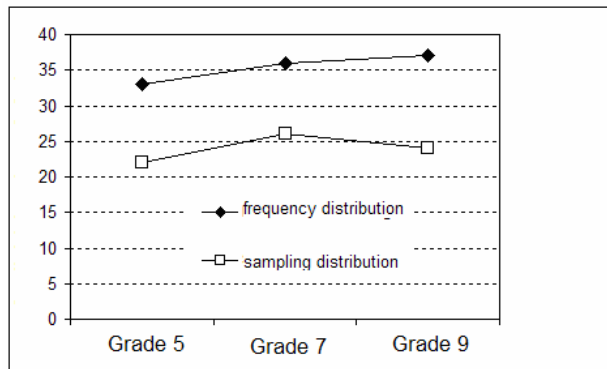
Figure 4. Summary of results for Task 2.



4.3 Influence of sample size

Figure 5 shows how sensitivity to the effect of the sample size in a variant of the maternity ward problem changes over the grade levels. For the aggregated data the advantage found earlier of the frequency distribution format is confirmed. However, the solution rates for both types of problem are lower than the average solution rate found among adult test persons (Sedlmeier & Gigerenzer, 1997). The comparatively low solution rate may hint at problems in understanding the task properly (Sedlmeier, 1998).

Figure 5. Results for Task 3. The vertical axis represents percentages of correctly solved subtasks.



Does the students' sensitivity to the effects of the sample size increase with the number of years of school attendance? Results as displayed in Figure 5 indicate that this, if at all, only happens to a very low degree. While with the frequency distribution format a slightly positive trend is discernible, this trend, initially also present in the sampling distribution format, turns around. One should also be aware here that the solution rates even in the frequency distribution format are almost at the chance level (there were only three answers to choose from!). Altogether, the sensitivity to the effect of the sample size on the precision of an estimate of a population parameter apparently does not improve with the increase in years of school attendance.

4.4 Probability of Conjunctive Events

In studies with adults the solutions rates in conjunction tasks of the type used in the present study are often not higher than 20%. Our results are at a similarly low level for the 9th graders, as shown in Figure 6. Furthermore, we notice a clear downhill trend. With increasing grade level the proportion of correctly solved subtasks falls. The temporal perspective, i.e. if the event to be judged is before or after the initial event (a D in the end-of-the-year report card) did not have any influence on the result. Surprisingly, the presentational format – frequencies versus probabilities – did not have a systematic effect on the solution rates either: while the frequency version tended to lead to slightly better results among 7th graders than the probability version, the conclusion for 5th graders is just the reverse. Results among the 9th graders differed by only about 1 percent in favor of the frequency format.

Figure 6. Results for Task 4. The vertical axis represents percentages of correctly solved subtasks.

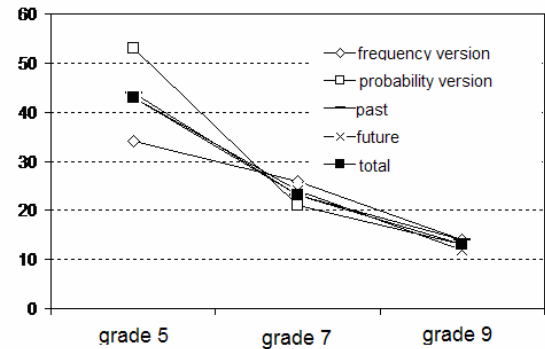


Figure 6 shows a clear trend: with increasing grade level, students were less capable of solving relatively simple problems involving conjunctive probabilities. The missing effect of a temporal perspective could mean either that this factor plays no important role when solving problems of this type or point to the fact that we did not succeed in simplifying the problem sufficiently. A possible reason for this could be that in all variants the events to be judged had already occurred (everything took place in the past). The missing effect of the presentational format was surprising, however. At the moment we cannot offer an explanation for that. The overall results, however, are clearly consistent with the assumption that statistical intuitions do not improve, but rather deteriorate, over the years of school attendance.

5. Implications for the Classroom

Our investigations show that students' comprehension of random events and chance variation in empirical data does not improve over the years of attending school during middle grades. On the contrary: if there is an observable trend at all, it points in the opposite direction of a deteriorating ability to deal with randomness and variability. This conclusion is consistent with earlier results obtained with comparable age groups by Green (1982) and generalizes the findings with an extended selection of tasks. To the best of our knowledge it is also the first study of its kind with German students. Which consequences are to be drawn from our results for the curriculum development of the middle school and for school policy?

Experiences with chance and randomness – similar to experiences with causal thinking – should be addressed in school on an interdisciplinary level. A variety of school subjects have to be involved in topics where probabilistic and deterministic views and explanations are juxtaposed and confront each other. As some examples, we suggest: in biology, connections between genetics and statistics; in natural sciences, weather forecasts and global warming; in economics, the stock market, as well as results of sports events or election polls. Experiments with random samples can be performed in social science studies as well. A special role, of course, is assigned to stochastics as the mathematics of chance. However, it seems very

unlikely that the kind of mechanical probability calculations asked for in many textbooks are of much help. These are often based on an *a priori* approach with the experimental side played down (so as not to spoil the pure mathematics). A stochastics classroom that stresses the aspect of formal computations without paying due attention to intuitions of statistical thinking will not do much good (see Rasfeld, 2004). Intuitive ideas and perceptions as the true sources of creative thinking (Borovcnik, 1992) are too strongly neglected then. Primary intuitions, i.e. ideas that evolve without a systematic treatment of a concept, are sometimes very useful in the classroom, at times they have to be questioned systematically. But they have to be addressed in school. Many misconceptions manifest themselves – as the plenty of paradoxes in stochastics show – if instruction neglects to build on early ideas (secondary intuitions) which require a systematic treatment and a connection between intuitions and a conceptual understanding based on the acquisition of a theory (see Tietze et al., 2002). Here, three factors seem to play a decisive role: the presentational format, i.e. how information is presented, “learning by doing“, and the proximity of the problems used to the students’ world (Sedlmeier, 1999). An instructional design that is dominated by a structural view of mathematics and that is less application-oriented in its treatment of stochastics would be quite counterproductive.

The acquisition of valid statistical (secondary-) intuitions can be realized as a gradual process only, for example by guiding learners gradually through experiments in games and simulation situations to a more formal way of representation (see e.g. Sedlmeier, 2001; Sedlmeier & Köhlers, 2000). Thus, we call for an instructional design that is characterized by students’ experimental activities including their own data collection and that takes advantage of new technologies (the computer, graphics calculator, the Internet), in order to plan and do simulations and demonstrations. Data produced repeatedly by a random number generator offer a substantial contribution to developing students’ intuitions about chance variation. By planning and carrying out experiments with random number generators and analyzing the results, students gain experience in stochastic situations and are enabled to check and – if needed – correct their intuitions about random events and probabilities.

The process of building appropriate intuitions has to begin with the students’ experiences and should be oriented towards practice and students’ daily life. A good starting point for introducing new concepts in stochastics instruction is student activities. This includes the planning and conducting of experiments, simulations and demonstrations as well as the construction of models and their interpretation. More specifically, in Engel (2002) a five-step procedure is suggested in the teaching of probability using simulation: 1) Introduction of a “real-world” problem involving some observations of reality or experiment; activity to experience the dynamics of the phenomena; 2) Building a simulation model based on a physical or software based random number generator;

3) Generating and analyzing data, including simulation-based inferences; 4) Critical evaluation of conclusions and reflecting on the impact of our assumptions by referring simulation results to the reality based problem we started with; and 5) analysis based e.g. on combinatorial arguments and formal probability depending on formal mathematical level of students.

Simulation is today a fundamental tool, that even in primary education allows students to model and experiment with random phenomena and predict the long run behavior, giving a reality to the frequency interpretation of probability. Simulation experiences, either through physical or computer based random number generators, provide a learning environment well suited to confront students with their erroneous beliefs and misconceptions about random phenomena (Batanero et al., 2005) and plays a very important role in developing valid stochastic intuitions. In simulating we substitute a real random situation by a different experiment, which is a model for the original but can easily be manipulated and analyzed. Between the domain of reality (the random situation we want to analyze) and the theoretical domain (a mathematical model) Coutinho (2001) locates the pseudo-concrete domain where we work with simulation. There, the student is out of reality and works with an abstract ideal situation. The didactical role of a pseudo-concrete model is to implicitly induce the theoretical model to the student, even when mathematical formalization is not possible (Henry, 1997).

However, we also realize that simulation can never replace thinking about a problem. Therefore, simulations alone can never directly attack wrong intuitions. They serve only as indication of where one’s intuitive thought is wrong. With the help of combinatorial schemes or tools like tree diagrams students start to understand the solution of probabilistic problems. This indicates the complementary nature of classical and frequentist approaches to probability. Therefore, we suggest that simulation is accompanied by formal mathematical analyses as well as by reflection on the pedagogical component of the situation. Misconceptions are not dissolved by just confronting learners with relevant situations which are prone to wrong approaches; instead of trying to revise wrong intuitions one should build a bypass by representing the situation in a more suitable format. The method and format of presenting information (frequencies or probabilities) may well facilitate the usage of more formal ways of presentation (Wassner, Martignon & Sedlmeier, 2002).

Randomness means that in a given situation there is more than just one possible result, that the realized outcome is not predictable with certainty, that – at least in our imagination – the possibility exists of repeating the experiment many times over and that the sequence of the repeated results has no pattern that can be controlled or predicted. Within this obvious chaos a number of regularities can be discovered, the most obvious of which is the stabilizing of relative frequencies of a possible event. This global regularity forms the foundation that allows us to study random

phenomena systematically according to the laws of probability. A basic understanding of randomness and probability is necessary for the understanding of the uncertainties of life and makes it easier to deal with these uncertainties. It would be a pity if this basic understanding and the theory that this understanding is based upon – probability theory – were to be withheld from our students as a result of inadequate scholastic instruction.

References:

- Batanero, C., Biehler, R. Engel, J., Godino, J., Maxara C. & Vogel, M. (2005): Using Simulation to Bridge Teachers' Content and Pedagogical Knowledge in Probability. *In press*
- Batanero, C., Godino, J., Vallecillos, A., Green, D. & Holmes, P. (1994): Errors and difficulties in understanding elementary statistical concepts. *Int. J. Math. Educ. Sci. Technol.*, 25 (4), 527-47.
- Borovcnik, M. (1992): *Stochastik im Wechselspiel von Intuitionen und Mathematik*. Bibliographisches Institut, Mannheim.
- Borovcnik, M. & Bentz, H. J. (1991): Empirical Research in Understanding Probability. In R. Kapadia & M. Borovcnik (eds.), *Chance Encounters: Probability in Education*. Dordrecht: Kluwer, 73-106
- Countinho, C. (2001). *Introduction aux situations aléatoires des le Collège: de la modélisation à la simulation d'expériences de Bernoulli dans l'environnement informatique Cabri-géomètre-II*. Unpublished Ph.D. University of Grenoble, France.
- Eichelsbacher, P. (2002): Mit Runs den Zufall besser verstehen. *Stochastik in der Schule* (22), Heft 1, 2-7.
- Engel, J. (2002). Activity-based statistics, computer simulation and formal mathematics. In B. Phillips (Ed.), *Proceedings of ICOTS-6*. CD-ROM.
- Fischbein, E. (1975): *The Intuitive Sources of Probabilistic Thinking in Children*. Reidel: Dordrecht-Holland.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, 103, 592-596.
- Green, D. R. (1982): A Survey of Probability Concepts in 3000 Students aged 11-16 Years. In D. R. Grey (ed.), *Proceedings of the First International Conference on Teaching Statistics*, Teaching Statistics Trust, University of Sheffield, 766-783.
- Green, D. R. (1986): Children's understanding of randomness. In R. Davidson & J. Swift (eds.), *Proceedings of the Second International Conference on Teaching Statistics*. Victoria, British Columbia, 287-291.
- Green, D. R. (1990): *A Longitudinal Study of Pupils' Probability Concepts*. Loughborough: Loughborough University.
- Henry, M. (1997). Notion de modèle et modélisation en l'enseignement. In *Enseigner les probabilités au lycée* (pp. 77-84). Reims; France: Commission Inter-IREM.
- Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning error. *Journal of Behavioral Decision Making*, 12, 275-306.
- Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. (E. A. Lunzer & D. Papert, Trans.). London: Routledge & Kegan Paul. (Original work published 1959).
- Kahneman, D. & Tversky, A. (1972): Subjective probability: a judgement of representativeness. *Cognitive Psychology*, 3, 430-454.
- Kahneman, D., Slovic, P. & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Kuzmak, S. D., & Gelman, R. (1986). Young children's understanding of random phenomena. *Child Development*, 57, 559-566.
- Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An Exercise in adversarial collaboration. *Psychological Science*, 12, 269-275.
- Nickerson, R. (2002): The Production and Perception of Randomness. *Psychological Review*, 109 (2), 330-357.
- Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children* (L. Leake, Jr., P. Burrell, & H. D. Fishbein, Trans.). New York: Norton. (Original work published 1951).
- Piaget, J. & Inhelder, B. (1975): *The Origin of the Idea of Chance in Children*. London: Routledge & Kegan Paul .
- Piattelli-Palmarini, M. (1994). *Inevitable illusions: How mistakes of reason rule our minds*. New York: Wiley.
- Rasfeld, P. (2004): Verbessert der Stochastikunterricht intuitives stochastisches Denken? Ergebnisse aus einer empirischen Studie. *Journal für Didaktik der Mathematik*, 25, 33-61.
- Sedlmeier, P. (1998): The distribution matters: two types of sample-size tasks. *Journal of Behavioral Decision Making*, 11, 281-301.
- Sedlmeier, P. (1999) *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah: Lawrence Erlbaum Associates.
- Sedlmeier, P. (2001). Statistik ohne Formeln. In M. Borovcnik, J. Engel & D. Wickmann (Hrsg), *Anregungen zum Stochastikunterricht*. Hildesheim: Franzbecker. (S. 83-95).
- Sedlmeier, P. (2002) Associative learning and frequency judgments: The PASS model. In P. Sedlmeier & T. Betsch (eds.), *Etc: Frequency processing and cognition*. Oxford: Oxford University Press (pp. 137-152).
- Sedlmeier, P. & Gigerenzer, G. (1997): Intuitions about sample size: the empirical law of large numbers. *Journal of Behavioral Decision Making*, 10, 33- 51.
- Sedlmeier, P. & Köhlers, D. (2001): *Wahrscheinlichkeiten im Alltag: Statistik ohne Formeln*. Braunschweig: Westermann.
- Shaughnessy, J. M. (1992): Research in probability and statistics: reflections and directions. In D. A. Grouws (ed.), *Handbook of Research on Mathematics Teaching and Learning*. New York: Macmillan. (S. 465-494).
- Steinbring, H. (1986): L'indépendance stochastique. *Recherches en Didactique des Mathématiques*, 7 (3), 99-118.
- Tietze, U., Klika, M. & Wolpers, H. (2002): *Mathematikunterricht in der Sekundarstufe II. Band 3: Didaktik der Stochastik*. Wiesbaden: Vieweg.
- Tversky, A. & Kahneman, D. (1983): Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, 90, 293-315.
- Wassner, C., Martignon, L., & Sedlmeier, P. (2002). Entscheidungsfindung unter Unsicherheit als fächerübergreifende Kompetenz. *Zeitschrift für Pädagogik*, 45, 35-50.

Authors

- Engel, Joachim, Prof., Faculty of Mathematics and Physics, University of Hannover, Im Welfengarten 1, 30167 Hannover, Germany
Email: engel@math.uni-hannover.de
- Sedlmeier, Peter, Prof., Department of Psychology, Technical University Chemnitz, Straße der Nationen 62, 09112 Chemnitz, Germany
Email: Peter.sedlmeier@phil.tu-chemnitz.de