Students' difficulties at the secondary-tertiary transition: the case of random variables

Camille Doukhan, Ghislaine Gueudet

Students' difficulties at the secondary-tertiary transition: the case of random variables

Camille Doukhan and Ghislaine Gueudet

CREAD, ESPE de Bretagne, University of Brest, France; camille.doukhan1@espe-bretagne.fr, ghislaine.gueudet@espe-bretagne.fr

The work presented here concerns the teaching and learning of random variables at the secondary-tertiary transition. Our study takes place in France, where discrete random variables are studied in grade 11 and in the first post-secondary year. Comparing textbooks, we observe that random variables are introduced as tools for the modelling of random situations at grade 11, while they are presented theoretically at the post-secondary level. The answers of first year students to a test confirm this difference: they are able to use random variables to produce interpretations in a concrete situation, but are not prepared for the learning of theoretical definitions. These results can contribute to design a teaching intervention concerning probabilities for first year students.

Keywords: Praxeologies, Probabilities, Random variables, Registers, Secondary-Tertiary transition.

Introduction

Difficulties encountered by students at the secondary-tertiary transition have been analyzed in many research studies in mathematics education, with different perspectives (Gueudet & Thomas, in press). Some of these studies focus on particular mathematical topics: calculus or linear algebra for example. The transition issues concerning probabilities have been less studied so far; yet probabilities are present in many tertiary courses, for non-specialists of mathematics in particular, like biologists or economists. Our study, taking place in France, addresses this issue with a focus on the concept of random variable. Our research questions are the following:

- How is the concept of random variable taught at secondary school, at the beginning of tertiary education? Which similarities or differences can be observed?
- What do the students know about random variables after the end of secondary school, and does this prepare them adequately for what is taught in the first year of tertiary level?

We study these questions with an institutional perspective. We present in the next section our theoretical tools, and studies about the teaching and learning of probability which inform our research. Then we expose in Section 3 the context of our study and our methods. In Section 4 we present our results, coming from a textbook study and a test proposed to first year students. We draw conclusions from these results and present them in Section 5.

Theoretical frame and related works

In our work we consider the secondary-tertiary transition as an institutional transition: the mathematics taught is shaped by the teaching institution, and is thus different at secondary school and university. We use the Anthropological Theory of Didactics (ATD, Chevallard, 2006), which introduces, in particular, the concept of praxeology. A praxeology has four components: a type of tasks, T; a technique, τ to accomplish this type of tasks; a technology θ which is a discourse

explaining and justifying the technique; and a theory Θ. Comparing the praxeologies at the end of secondary school and at the beginning of university can illustrate different kinds of changes in the mathematics taught (see e.g., Winsløw, Barquero, De Vleeschouwer & Hardy, 2014). Some changes are very general: for a given type of tasks, a single technique is taught at secondary school, while several techniques can be expected at university; similarly, at university the theoretical block (formed by the technology and the theory [θ,Θ]) is more important than the practical block [T, τ]. Other changes concern a particular type of task, for example in our case: "Determine the probability distribution of a random variable," studied below.

We also use in our work the concept of semiotic registers (Duval, 2017). Mathematical concepts, in particular random variables, can be described through different semiotic registers: a text, a table, or a graph. Compared with secondary school, the institutional expectations concerning the flexible use of different registers increase at university (Gueudet & Thomas, 2018). Hence observing the registers used, in particular the need to combine several registers in the same technique for a given task, can illustrate discontinuities in the transition process.

Concerning probability, several works acknowledge the specific nature of probabilistic reasoning. Batanero et al. (2016) state for example that "probabilistic reasoning is a mode of reasoning that refers to judgments and decision-making under uncertainty and is relevant to real life" (p. 9). Such link between probability and real-life is emphasized by many authors, and in the curriculum of different countries (Batanero et al., 2016). However, some studies indicate that an important proportion of probability problems proposed to students are not related to real life situations and that the problems involve more algebraic than probabilistic reasoning (20% in entrance tests to university in Andalucía according to Batanero, López-Martín, Arteaga, & Gea, 2018). Student difficulties are observed and connected with the teaching practices; for example Burrill and Biehler (2011) claim that a formal teaching of probabilities focusing on set theory and counting techniques can hinder the understanding of the associated phenomena.

Research in mathematics education concerning random variables is scarce. At the beginning of university, Amrani and Zaki (2015) observed student difficulties, after a very formal teaching about random variables. Given a random experiment, they had difficulties proposing a model with an appropriate random variable and determining the probability distribution for a given random variable. Nevertheless these authors did not observe the impact of a less formal teaching.

Context and methods

Our research takes place in France, where probability is an important part of the official curriculum at secondary school, starting at Grade 9. We focus here on Grade 11 of the scientific track (called "S section"). Students encounter random variables for the first time at Grade 11, in the context of discrete random variables. The concepts of *probability distribution*, *expected value*, and *variance* are defined; the Bernoulli and binomial distributions are introduced. According to the official curriculum: "*The concept of probability distribution of a random variable permits the modelling of random situations and their study with the tools of probability*". At Grade 12 students encounter continuous random variables, with the concept of density and the Gaussian distribution.

We chose to focus on "preparatory classes" in the tertiary level, selecting high-achieving students, who want to enter a "*grande école*" after two years of preparation. The classes are organized is similarly to secondary school, hence we claim that the institutional differences mainly come from the mathematics taught. Moreover, because they are high achieving, the difficulties encountered by these students should not come from an insufficient grasp of secondary school praxeologies. The official curriculum of the first year of preparatory classes (in our case an "Economics and Commercial –Scientific, ECS," preparatory class) concerning random variables is limited to discrete random variables (continuous random variables appear only in the second year). The curriculum revisits the notions taught at Grade 11: random variable, probability distribution, expected value, and introduces the study of pairs of random variables, and independent random variables. The official curriculum states that "Modelling simple random situations with random variables is a skill expected from students", meaning that the preparatory classes institution is not responsible for this modelling skill.

In the study presented here, we used two kinds of methods. First, we performed a comparative analysis of a secondary school textbook, and a preparatory class textbook. The secondary school textbook (Barbazo, 2015), is used by a large number of Grade 11 teachers in France. The students in the preparatory classes did have a textbook, but they often bought books to complement the course given by their teacher. The content across books was quite similar, so we chose Gautier and Warusfel (2012). We compared the chapters concerning random variables according to five characteristics: the structure of the chapters in terms of course vs exercises (theoretical vs practical block), the definition of random variable given (considered as an important aspect of the theory), the registers of representations used, the types of tasks and associated praxeologies, and the contexts in which the exercises and examples were situated. For each chapter studied, we listed and counted the types of tasks requested in each exercise or worked example. The theoretical block appeared mainly in the content parts of the chapters (definitions, theorems, demonstrations, proposals) whereas the practical block appeared in the exercise and example parts of these chapters.

Second, we designed a test and submitted it to the students from a first year ECS class before the teaching of random variables (see Appendix). The teacher only allowed us to use 20 minutes to administer the test in class, so we had to limit its length. The test started with three questions: mathematical definition of a random variable, how to explain to a Grade 12 student what a random variable is, and give an example of random variable; then an exercise in an economical context was proposed. This exercise, taken from a website for Grade 11, was typical from what can be found in secondary school textbooks for this level. It corresponded to a random situation in a "real" context. Nevertheless the students' responsibility for modelling was reduced: the random variable to be studied was given in the first question. The types of task were: "Determine the distribution of a random variable (described by a text)"; "Compute the expected value of this random variable and interpret it"; and "Study the influence of a change in the initial data on the expected value". Forty-four students took the test at the beginning of a course. We analyzed the students' answers using the following categories: for Questions 1 and 2: elements of theory and technology used, errors, ideas or elements related to randomness, and type of answer (theoretical, related to randomness, other). For the Question 3 we looked at example correct or not, if not why, in what context it was stated,

and which registers were used. For the exercise, the categories were: answer correct or not, the technique used, the technology (justification of the answer), use of registers, and result correctly interpreted.

Modelling "real" situations and mastering formal definitions

Comparing secondary school and preparatory classes textbooks

We noted a large difference in the proportions of courses/exercises within the chapters studied, in terms of number of pages dedicated to each: 15% for course in the Grade 11 textbook; more than 80% in the ECS textbook. We consider this as an indicator of the much greater importance of the technological-theoretical block in higher education, compared to secondary education.

Concerning the definition of random variable, in the Grade 11 textbook the students already knew the concept of sample space Ω; a random variable X was defined as "a function from Ω to *IR"*. We claim that this definition was difficult for Grade 11 students, who mostly met functions from *IR* to *IR*. The definition of the probability distribution; as a function from *IR* to *IR*, followed immediately and it was more familiar. In the ECS textbook, the definition of random variable was situated at a much more general level. The definition of σ–algebra was introduced together with the notion of probability space as a triple (Ω, $\mathcal{F}$, P). In this context a very general definition of a random variable was given: "an application X from Ω to IR such that $\forall x \in$ IR, $\{\omega \in \Omega / X(\omega) \leq x\}$ belongs to $\mathcal{F}$".

We categorized the registers of representations present in the parts of courses, in the examples of the courses and in the exercises of the chapters studied. We observed seven different registers: symbolic, natural language, and five kinds of graphical registers: tree diagrams, function graphs, tables for representing the random variable distribution, some charts representing the random situation and various illustrations (e.g., photos). The symbolic register and the register of natural language were omnipresent in both textbooks. The graphical registers were more frequent in Grade 11 (68 examples used these registers) than in ECS (18 examples). There were, in particular, many illustrations in Grade 11, and none in ECS. There were 36 tables representing the random variable distribution in Garde 11 and only six in ECS. Overall, there were very few tree diagrams. In the secondary school textbook, random situations were often represented by charts or diagrams, but relatively few by trees.

Regarding the contexts in which the examples and exercises were posed, we distinguish two categories: exercises and examples in a non-mathematical context and in a mathematical theoretical context (what Batanero, *et al*. 2018 call "problems with no context," p. 113). This choice is governed by the claim of the secondary school curriculum, about the importance of random variables for "the modelling of random situations" (Ministère de l'Education Nationale, 2010, p. 5).

In the first category, *exercises and examples in a non-mathematical context,* we identified real-life situations such as the lifespan of a robot or the waiting time at a doctor, but also "artificial" situations such as a flea moving on a graduated axis or a drawing of balls from an urn. There were also situations in other scientific fields such as the study of physical uncertainty of measurement. The proportion of examples and exercises in a non-mathematical context is 78% for Grade 11 and 67% for ECS (out of a total of 108 exercises and examples, respectively 66).

In spite of this difference, we observed similarities concerning the types of tasks. For both Grade 11 and ECS, the three most frequent types of tasks were: “determining the probability distribution of a random variable,” “proving that a random variable admits an expected value and/or calculating it,” and “proving that a random variable admits a variance and/or calculating it.”

For example, for the type of task “determining the probability distribution of a random variable,” the following technique was described in the Grade 11 textbook: “(1) determine the possible values x_i for X; (2) compute the probabilities P ($X = x_i$); (3) summarize them in a table”. This technique was linked with the fact that in most cases, the random variable took a limited number of values (in our test for example, X takes 4 different values), that could be presented in a table. In the Barbazo textbook, 104 exercises amongst 106 were about random variables taking a limited number of values. In the ECS textbook, in 20 exercises amongst 21 the random variable took its values between 1 and an integer n. To determine the distribution, the technique is consisted of finding a formula giving P($X=k$) for $1 \leq k \leq n$. The associated technologies in the two institutions were provided by the definitions mentioned above. Thus the praxeologies were clearly different in each textbook.

Also the differences concerning the context lead, for similar types of tasks, to different responsibilities of the students in terms of modelling. In Grade 11, 19 exercises asked students to calculate an average payout or determine whether a game would be favorable or not to the player, with no intermediate tasks. For this type of task, the technique used was to select the appropriate random variable, and then determine its distribution in order to calculate the probability of a certain event. This modelling work is totally absent for ECS, although the proportion of exercises in the context of artificial situations is high. In these exercises, the random variable is always given and the types of tasks present are "determining the probability distribution" or "calculating its expected value".

Students responses to the test

We recall here that the test started with three questions, concerning: the mathematical definition of a random variable, the explanation for a Grade 12 student of what a random variable is, and an example of random variable. These questions were directly linked with the theoretical block [θ,Θ]. Twenty-three of 44 students (52%) answered the first question, 33 of 44 students (75%) answered the second question, and 26 of 44 students (59%) answered the third question.

The answers were very often false or incomplete. The most common errors were confusion between the random variable and the probability distribution (which we expected, considering the definitions given in the textbook) and confusion between the definition of a random variable and its properties. In the first question concerning the theory, six students confused the definition of a random variable with its properties, and five students confused the random variable with the probability distribution. In the second question (in which students are expected to produce a technological discourse), three students confused the random variable with the probability distribution. In the third question, 11 students confused the random variable with the probability distribution.

The vocabulary and mathematical notations used by the students allowed us to distinguish two types of answers: theoretical answers in which we find a symbolic register with mathematical

concepts and notations, and answers in relation to randomness in which we find a vocabulary in relation to randomness. Such answers contain ideas linked to chance, for example: “randomly draw a sample,” “results of a random experiment,” or “depends on chance.”

For the first question, students mainly gave a theoretical answer (19 of 23 students); only 4 gave an answer in relation to randomness. An example of theoretical response by student S2 is:

> “A random variable is a law that associates a variable entity (price, size) with X.” (our translation)

The second question generated more diversity in the registers used by the students in their responses. We distinguish for this question a new type of response, in which students responded with a specific example of the use of a random variable. Almost a fourth of the students produced such a technological discourse, explaining in their answers how a random variable can be used (see the example below). These students seem to have incorporated the perspective of the official curriculum: random variables as tools for the modelling of random situations.

An example of this type of response by student S38 is as follows:

> "it will be a question of insisting on the fact that a random variable is not fixed, it will allow specially to make calculations of returns, of profits. Know through expected value if it is relevant or not to play a game of chance. In order to be able to work on its expected value for example it will be necessary to draw up the law of probability (table of values)". (our translation)

Finally, on the third question, we were interested in the contexts in which students set their example. Eight of 26 students who answered this question gave an example in a real-life situation (e.g., the price of a chocolate bar, the price of a car, the life of a light bulb or the size of a plant) while four of 26 of students gave an example in an artificial context (e.g., a betting game, a die roll, or a drawing of balls from urns).

The students were more successful with the exercise than with the three preliminary questions. Thirty-seven of 44 students answered the first question that asked them to determine the distribution of a random variable (described by a text), 38 of 44 students answered the second question that asked them to compute the expected value of this random variable and interpret it, and 29 of 44 students answered the third question that asked them to study the influence of a change in the initial data on the expected value. Students who responded to these questions often offered a correct and well-justified answer.

Concerning the first question, a fourth of the students who answered used a technique involving the representation of the probabilistic situation with a diagram or a tree diagram. Questions of interpretation were often well addressed. Almost two thirds of the students who answered the second question gave a correct interpretation of the expected value.

Finally, we note that these students (who were high-achieving at secondary school) were able to use a random variable for the modelling of a concrete random situation. They interpreted the situation to find the distribution of the random variable, used it to solve a question concerning an expected value, but were unable to produce the correct theoretical/technological discourse needed at tertiary level.

Conclusion

The questions we studied here about the concept of random variable concerned on the one hand the differences between secondary and tertiary levels; and on the other hand how well prepared students were by their secondary school for the teaching proposed at tertiary level.

The textbooks' comparative study evidenced as expected that the practical block was more important at secondary school while the theoretical block was more important in the preparatory classes. The definition of random variable given at secondary school was less general, and immediately associated with the definition of the distribution, which is central in the exercises. Similar types of tasks were found in preparatory classes, but they corresponded to different praxeologies. For example, techniques were different because random variables met at secondary school had a limited number of values, while random variables in preparatory classes had values between 1 and *n.* At secondary school, random variables are introduced as tools to study situations and the focus was on the distribution, represented by a table. In preparatory classes, in contrast, the concept of random variable and the associated theory became central.

The student responses to the test are coherent with the textbook study: they were able to use a random variable for the study of a given situation but did not seem to handle well the theoretical and technological aspects. In particular they confounded random variable with its distribution. The textbook analysis suggests that secondary school teaching is probably not a sufficient preparation for what is expected at tertiary level, and our small sample of students seems to confirm this. Even if the concepts are the same, students at tertiary level are likely to meet difficulties for connecting the new theoretical aspects with their previous knowledge.

Beyond these observations, similar to the results obtained by studies about the secondary-tertiary transition for other topics (Gueudet & Thomas 2018), our work raises theoretical and methodological issues. Concerning the theory, the importance of modelling within the theme of probabilities leads to questions about "modelling praxeologies": how can the praxeology be described? Which kind of theory is involved in it? It is necessary to answer these questions, in particular, to design a teaching intervention in which modelling plays a central role, and theory is also present. Concerning the methods, we have observed that the analysis of textbooks can contribute to identify changes of praxeologies between different institutions. Nevertheless, identifying the theoretical and the practical blocks in a textbook is not always straightforward; different textbooks can have very different structures. We consider that this is a direction open for a further methodological work, especially important for transition studies.

Naturally the study presented here remains limited. In a further work we will consider the learning of probabilities by non-specialists in the first year of university, with the aim to propose a teaching and associated resources combining the modelling of real-life situations and theoretical aspects.

References

Amrani, H., & Zaki, M. (2015). Student's conceptual difficulties with respect to the notion of random variable. *International Journal of Education, Learning and Development*, *3*(9), 82-98.

Barbazo, E. (Ed.). (2015). *Mathématiques Barbazo 1re S*. Paris: Hachette.

Batanero, C., Chernoff, E. J., Engel, J., Lee, H., & Sánchez, E. (2016). *Research on teaching and learning probability*. ICME13 topical series. New York, NY: Springer.

Batanero, C., López-Martín, M., Arteaga, P., & Gea, M. M. (2018). Characterizing probability problems posed in university entrance tests in Andalucía. In C. Batanero & E. J. Chernoff (Eds.), *Teaching and learning stochastics* (pp. 103-123). New York, NY: Springer.

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* (pp. 57–69). Dordrecht, The Netherlands: Springer.

Chevallard, Y. (2006). Steps towards a new epistemology in mathematics education. In M. Bosch (Ed.), *Proceedings of the Fourth Conference of the European Society for Research in Mathematics Education* (pp 22–30). Barcelona, Spain: Universitat Ramon Llull Editions.

Duval, R. (2017). *Understanding the mathematical way of thinking – the registers of semiotic representations*. New York, NY: Springer.

Gautier, C., & Warusfel, A. (2012). *J'intègre, Mathématiques ECS 1ère année*. Paris: Dunod.

Gueudet, G., & Thomas, M. (in press). *Secondary-tertiary transition in mathematics education.* Encyclopedia of mathematics education. New York, NY: Springer.

Ministère de l'Education Nationale (2010). *Bulletin officiel spécial n°30. Mathématiques. Cycle terminal de la série scientifique. Classe de Première*. Retrieved from: https://cache.media.education.gouv.fr/file/special_9/21/1/mathsS_155211.pdf

Winsløw, C., Barquero, B., De Vleeschouwer, M., & Hardy, N. (2014). An institutional approach to university mathematics education: From dual vector spaces to questioning the world. *Research in Mathematics Education, 16*(2), 95-111.

Appendix: English version of the test given to the students

Please answer this little test without consulting any document: the goal is to know what you remember, at the end of high school, about random variables.

1. Give the mathematical definition of a random variable.
2. How would you explain what a random variable is to a senior student?
3. Give an example of a random variable.
4. Solve (on the back) the exercise below.

Optimization of a profit:

The cost of production of a laptop computer by the company MARBIC is 950€.
This laptop may have a defect A, a defectt B, or both the defect A and the defect B.
The warranty allows repairs to be made at the manufacturer's expense with the following costs: 100€ for the defect A, 150€ for the defect B et 200€ for the defect A et B.
We admit that 90% of the laptops produced have no defects, 4% have the only defect A, 2% have the only defect B and 4% have the defects A and B.

1. We denote X the random variable which, with each laptop randomly chosen, associates its cost price, i.e. its production cost increased by the possible repair cost.
   Determine the probability distribution of X.
2. Calculate the expected value $E(X)$ of this random variable. What does $E(X)$ represent for MARBIC? It's assumed that all laptops produced are sold.
3. (a) Can MARBIC hope to make a profit by selling 960€ each product?
   (b) MARBIC wants to make an average profit of 100€per laptop. Explain how to choose the selling price of the laptop produced.