ALIGNING EVALUATION WITH ACHIEVEMENT OBJECTIVES: AUTOMATED EXAMS BASED ON BLOOM'S TAXONOMY

Eduardo León Bologna, Marcelo Vaiman, Matías Adrián Alfonso
National University of Córdoba, Argentina
eduardo.leon.bologna@unc.edu.ar

How many of social sciences students passing introductory statistics courses develop the expected skills to make a meaningful use of statistics? Our diagnosis suggests that an important part of them achieve this through memorization and repetition. This communication reports the in-progress effort to improve the quality of the evaluation of an introductory statistics course in Psychology degree, National University of Córdoba (Argentina). There is a specific demand on the qualifications required of students who pass the subject, which combines with a significant volume of students, so it is necessary to ensure the validity of the evaluations and the automation of their administration and correction. The work consists of the construction of examination items classified according to three criteria: elementary thematic unit it evaluates, cognitive level and degree of difficulty, so that precision exams can be built. The proposal is applicable to classroom or on-line courses.

INTRODUCTION

Each year many Psychology students, at National University of Córdoba (Argentina), pass the first year introductory course called "Psychostatistics"; an annual compulsory subject that receives approximately 1500 students a year and meets a support role to directly correlated subjects, such as Psychometric Techniques and Research Methodology (in second and third year respectively). It is placed in the first year because it is necessary for the subsequent courses, and because it provides tools for subjects that are part of the experimental block and neurosciences, in which it is necessary to interpret results of quantitative research. The demand on the subject is the understanding of results expressed in statistical language, the possibility of transferring the forms of statistical reasoning to problematic situations, the identification of the limitations of each statistical technique and the necessary safeguards for its application and interpretation. In this sense, the objectives of the course of Psychostatistics are based in the four pillars of learning for the 21st century of the Delors´s report for the UNESCO (Delors, 1996, Burnett, 2008): learning to know, learning to do, learning to be and learning to live together. In relation to the first of them, the goal is to provide the students with a basic knowledge of the discipline that allows him to expand knowledge according to the needs that arise throughout his scientific and professional practice: knowing techniques of organization, summary and treatment of data. In relation to the second pillar, the course is oriented to student develop a competency that allows apply statistical tools to Psychology research problems. Regarding the third pillar, the objective is aimed at developing critical judgment, creativity, the ability to communicate with others and statistical thinking (Chance, 2002). In relation to the last pillar, the course brings the possibility to know and work with others: team working is needed to resolve activities.

However, despite the proposed framework, and that every year, approximately 80% pass this course, among them, the learning show insufficient: the year after, students are unable to handle the tools required to perform psychometric analysis, and two years later they cannot compare two experimental groups. The explanation to this situation is complex. On the one hand Psychostatistics is difficult because psychology students arrive with little or no interest in highly structured content, such as mathematics and statistics. In addition, schools from which they come from are very diverse in the depth they teach these contents, to the point that in some schools, statistics is completely absent.

The main problem that social sciences´ students face when trying to learn statistical concepts is the belief about their own inability. The frustrating experiences they may have had in school with mathematics, possibly influenced the decision to pursue a career in that field, and those experiences prevent the attempt to understand, handle, and make use of quantitative data. Ruggeri et al. (2008) remark that Psychology, as well as other social sciences, is often chosen by those students who show less interest and more negative self-assessments in mathematics, physics, etc., who also use to underestimate the extent to which statistics is present in the subjects they choose. For many students there is a barrier prior to the attempt to understand, founded in the scarce self-confidence to learn

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. iase-web.org/Conference_Proceedings.php

contents that evoke mathematics. The idea of an innate inability for mathematics radiates towards statistics and limits the time and effort the student devotes to trying to understand.

This incapacity is well described by the concept of self-efficacy in the sense of Bandura (1997), which indicates the achievement the individual perceives himself capable of reach (Bandura, 2012) and that, due to its specificity, distinguishes from other constructs such as self-esteem or self-concept (Bandura, 1997) and, according to Pajares & Miller (1994), is a better predictor of achievement than self-concept or perceived utility.

It also contributes to the difficulty of studying statistics, the anxiety it produces (Zeidner, 1991), especially in degrees such as Psychology, Education, or Sociology (Onwuegbuzie & Wilson, 2003). As a consequence, statistics anxiety is supposed to lead to manifold problems over the course of students' statistics education. Students who experience higher levels of anxiety are expected to be more likely to postpone doing application homework, to study for examinations, or to keep up with the readings (Onwuegbuzie & Wilson, 2003). Given the need to pass the examination of Psychostatistics to advance in the studies, added to the anxiety that supposes to feel incapable of understanding the themes, the students look for strategies that provide shortcuts (Ortega, 1996, 2008) to overcome the obstacles in the shortest time and with the least effort. Many of them are oriented to learn what is necessary to pass, a strategy that is usually accompanied by the memorization based on repetition.

Reflecting on this, we consider that these strategies are in turn reinforced by the evaluation system. Taking into account what Ortega (1996) says, that they are likely to study to pass, it is expected that they study to respond to certain forms of evaluation. In this sense it is necessary that the teaching process contemplates the effect that the way of evaluating has on the way that students learn. One way to approach to this problem is take the contributions of Bloom's Taxonomy of Educational Objectives (1956) and its subsequent revisions (Bloom, Engelhart, Hill, Furst, & Krathwohl, 1956; Krathwohl, 2002). This classification system allows organizing learning and evaluation objectives at different levels of complexity. To do this, the taxonomy distinguishes a hierarchy of six cognitive skills: knowledge, understanding, application, analysis, synthesis and evaluation. These skills range from a minimum that is the possibility of evoking previously learned information (facts, concepts and procedures), to the possibility of making judgments and criticisms based on given criteria. This frame gives an operational definition of what students who pass the exam are expected to achieve.

However, since students doubt their ability to understand the contents of the subject, they choose the shortcut of studying them in a memoristic way, to make mechanical applications and to apply the procedures in a routine way without understanding them. If the evaluations are not well designed, this can be a successful strategy. This way, the student could reach his immediate objective of approving the subject, without having met the learning objectives. When rote learning is enough to pass the exams, the evaluation could even limit meaningful learning. How can we manage to avoid these shortcuts and maximize to probability of meaningful learning for those who pass the exam? It is not our aim to build more difficult exams, but to guide students towards the development of learning skills, since the exam is not only intended to assess student progress, but also to facilitate learning opportunities (Bush, Daddysman, & Charnigo, 2014).

STRATEGIES

In order to reach the achievement expectations deposited in those who pass the exam, considering the limitation imposed by massive courses, for almost a decade we have been modifying the way of dictation in search of establishing connection between the contents offered in each thematic unit and to link these contents with specific problems of Psychology. A book (Bologna, 2014, 2018) was prepared with examples taken from research carried out in the faculty itself to make more sense of the subject. The practical activities establish these links, because the students work all year with data collected by them interpreting results and transforming in text the statistical outputs from software. In addition, in order to reduce the effect of the prejudices, we have outlined pedagogical strategies that seek a gradual approach to content and a gradual strengthening of the confidence that students gain in their abilities. One action of this strategy is that the first partial exam is of low difficulty, in order to reduce the threatening character of the subject and to value the first achievements in the appropriation of the contents (Bologna & Vaiman, 2013).

We now are facing the problem of achieving assessments that oblige students bring into play cognitive levels superior to the mere recognition of concepts and techniques. It is a way to achieve

meaningful learning, since the conceptual evaluation does not ensure that students have gone beyond having memorized certain routines for reading measures and indicators.

In multiple choice exams there seems to be an inverse relationship between the advantage that students can obtain from rote learning and the diversity of questions that may result in exam. The uncertainty about the exam questions obliges students to develop adaptive strategies, which must be flexible enough to cope with problems presented in varied ways. When it is difficult to guess the questions that may appear, the most efficient strategy is to master the subject.

A good quality test is the one that lets you know that a high proportion of those who pass it reach the objectives of the subject. This requires an exam that covers all the contents of the program, that evaluates the cognitive domains that are required, and that is unpredictable, in order to discourage memoristic study. Considering the high volume of students, it must also be standardized. This demands a broad question bank and the personalized generation of a set of items for each student at the moment of evaluation.

In addition, the degree of difficulty of each item is a valuable piece of information to graduate the exams, to identify the groups of students in different levels and to detect the harder contents. The relative difficulty of themes may vary from one course to another, due to different characteristics of the students and different emphasis teacher can place on each theme. The knowledge of the degree of difficulty of different contents allows orientating the dictation forms and practical activities.

METHODS

The first stage to build a good quality assessment instrument was the fragmentation of the contents of the syllabus into elementary units. Beginning with "definition of variable" and arriving to "t tests", 59 micro-contents were identified (such as: "skewness and mean interpretation", "effect of sample size on confidence intervals"). These contents were then crossed by Bloom's cognitive levels to give rise to a 59 X 6 matrix that classifies the questions according to the two categories: content and cognitive level, resulting in 354 question categories. The question bank is constructed by generating one or several question structures for each cell of the matrix, to evaluate the corresponding content in the corresponding domain. Here is an example:

"With a confidence of (1-α), we estimate the reaction time to a stimulus in [Ul; Ll]. If the confidence changes to (1-α'), keeping everything else fixed, which of the following intervals could correspond to the estimation of the same parameter ". K answer options are offered, one of which is correct; the one that centers the interval in the same value and adjust the error according to α and α'. Here the evaluated micro-content is *effect of the confidence in the estimation error* in the domain *analysis* according to Bloom Taxonomy. Then, the question bank is generated reproducing the structure and modifying the values of α, α', Ul and Ll in a random way. These operations are performed by means of an R (R Team Core, 2015) package, called exams (Zeileis, Umlauf, & Leisch, 2014), which uses Rmarkdown (Xie, Allaire, & Grolemund, 2018) syntax. The set of items thus generated is exported to Moodle for the development of models with N questions randomly chosen from that base.

For the generation of the exam items and their export to the Moodle platform, a public project has been created at https://github.com/mentoldo/exam_stat/tree/development (Alfonso & Bologna, 2018), with the code for the automatic generation of 45 questions. In the question bank of Moodle, it is important to be careful of the way in which the items are organized to enable the construction of the exam models. A parent category (in the Moodle language) will be Confidence Intervals and, within it, there are subcategories that represent variations around a thematic micro-content. For the present example, that subcategory is called Intervals. Confidence effect, and contains 100 variations around the structure showed in the former example.

For the empirical evaluation of the difficulty level of each question we make use of the measures that Moodle provides to analyze the items of a questionnaire, specifically the Facility Index. This measure indicates the proportion of times the question was answered correctly with respect to the total number of times it was administered. This index can be requested on the set of questions of the parent category or on those that make up each subcategory within it. The quality of the questions can also be assessed by mean of other measures Moodle offers, what allows monitoring and adjustment in a continuous improvement process.

CONCLUSION

The proposal here exposed is in progress, it takes time to develop structures of questions that match simultaneously the micro-content and the cognitive level. This proposal combines four elements: i) a detailed thematic classification of the contents of the subject, which identifies elementary units; ii) their distribution according to the cognitive level to which they refer, following Blooms taxonomy; iii) the production of a large number of evaluation items that inquire about the same content, with variations in the numerical values and in the response alternatives; and iv) the empirical analysis of the difficulty that each question generates in the students to keep track of the themes that appear harder to learn for them.

With this procedure, we believe that it is possible, in a gradual way, to obtain valid evaluations, that is, to give the highest degree of certainty that their approval implies having reached an acceptable cognitive level on the subjects that are evaluated. It also allows the monitoring of learning in an aggregate level, as a feedback to improve teaching strategies. At individual level, it serves to identify students who need more attention. And this is done in the framework of the administration of exams to large volumes of students, so it appears as applicable to both face-to-face and on-line courses.

REFERENCES

Alfonso, M., & Bologna, E. (2018). Generación automática de preguntas de estadística (Version 1.0). https://doi.org/10.5281/zenodo.1465036

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York, NY: W H Freeman/Times Books/ Henry Holt & Co.

Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited. *Journal of Management*, *38*(1), 9–44. https://doi.org/10.1177/0149206311410606

Bloom, B., Engelhart, M., Hill, W., Furst, E., & Krathwohl, D. (1956). *The taxonomy of educational objectives.* New York, NY: McKay.

Bologna, E. (2018). Métodos Estadísticos de Investigación (1st ed.). Córdoba, Argentina: Brujas.

Bologna, E. (2014). Estadística para Psicología y Educación. (Brujas, Ed.) (3rd ed.). Córdoba, Argentina: Brujas.

Bologna, E. L., & Vaiman, M. (2013). Actitudes, experiencia previa y nivel de logro en Estadística en la carrera de Psicología. Probabilidad Condicionada: Revista de didáctica de la Estadística, (1), 91–104. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=4770239&info=resumen&idioma=SPA

Burnett, N. (2008). The Delors Report: a guide towards education for all. *European Journal of Education*, *43*(2), 181–187. https://doi.org/10.1111/j.1465-3435.2008.00347.x

Bush, H. M., Daddysman, J., & Charnigo, R. (2014). Improving Outcomes with Bloom's Taxonomy: From Statistics Education to Research Partnerships. *Journal of Biometrics & Biostatistics*, *5*(4), 4–6. https://doi.org/10.472/2155-6180.1000e130

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, *10*(3).

Delors, J. (1996). Los cuatro pilares de la educación. En: La educación encierra un tesoro. Informe a la UNESCO de la Comisión internacional sobre la educación para el siglo XXI, Madrid, España: Santillana/UNESCO. pp. 91–103

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, *41*(4), 212–218. Retrieved from https://www.depauw.edu/files/resources/krathwohl.pdf

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments - a comprehensive review of the literature. *Teaching in Higher Education*, *8*(2), 195–209. https://doi.org/10.1080/1356251032000052447

Ortega, F. (1996). Docencia y evasión del conocimiento. *Estudios: Centro d Estudios Avanzados*, (7), 5–15.

Ortega, F. (2008). Atajos. Saberes escolares y estrategias de evasión. *Bs. As. Miño y Dávila Eds*.

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, *86*(2), 193–203. https://doi.org/10.1037/0022-0663.86.2.193

R Team Core. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Ruggeri, K., Díaz, C., Kelley, K., Papousek, I., Dempster, M., & Hanna, D. (2008). International Issues in Education. *Psychology Teaching Review*, *14*(2), 65-74.

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. CRC Press. Retrieved from https://bookdown.org/yihui/rmarkdown/

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology*, *61*(3), 319–328. https://doi.org/10.1111/j.2044-8279.1991.tb00989.x

Zeileis, A., Umlauf, N., & Leisch, F. (2014). Flexible Generation of E-Learning Exams in R: Moodle Quizzes, OLAT Assessments, and Beyond. *Journal of Statistical Software*, *58*(i01). Retrieved from http://cran.r-project.