STATISTICS TEST QUESTIONS: CONTENT AND TRENDS

AUDY SALCEDO
Universidad Central de Venezuela
audy.salcedo@ucv.ve

ABSTRACT

This study presents the results of the analysis of a group of teacher-made test questions for statistics courses at the university level. Teachers were asked to submit tests they had used in their previous two semesters. Ninety-seven tests containing 978 questions were gathered and classified according to the SOLO taxonomy (Biggs & Collis, 1982) and to the definitions of statistical literacy, statistical reasoning and statistical thinking (delMas, Ooms, Garfield & Chance, 2007). Results suggest a strong preference for questions that address the evaluation of cognitive abilities in the lower levels of the taxonomies used. Reflections as to the implications of these results for the teaching and evaluation of statistics courses are presented.

Key words: *Statistics education research; SOLO taxonomy; Statistics literacy; Statistical reasoning; Statistical thinking*

1. INTRODUCTION

Within formal education systems the term 'evaluation' refers to at least three important processes: the instruction conducted by the teacher, the curriculum, and student achievement. However, evaluation of student achievement is considered – without underestimating the other two processes – particularly relevant because it has to do with a fundamental aspect of the educational process: the outcomes in terms of the relationship between teaching and learning. Despite the fact that it is often recommended that various and varied means of evaluation be used, teacher-made tests continue to be the preferred means of student evaluation in the system of higher education in Venezuela. Even when other means of evaluation may be used, tests usually tend to play a major role in the gradings. Considering this, written and oral exams play a very important role in making decisions as to the promotion and accreditation of students in tertiary education, especially in those courses in the quantitative area where by tradition the written test is the customary means of evaluation. Consequently, this research addresses a particular aspect of written statistics exams as it seeks to analyze questions used to evaluate student learning in the context of the subject area of Statistics Applied to Education.

Exams are usually designed by the teacher who, as an expert in statistical content, makes decisions in terms of what kind of assessment is to be used, which aspects of the program are to be evaluated via written tests and which others by different means. When designing an exam, the instructor selects the items to be used in order to gain information related to the progress made by students during course time. This process is carried out according to course objectives set, but also having in mind other aspects such as content previously addressed, ongoing learning experiences or student characteristics. In relation to this, test design is not a process in which teachers make use of technical considerations only, but one that also involves their personal ideas about teaching, learning, evaluation, and about the discipline they teach.

Related to the above, Shulman (1986) has highlighted the need to take into account the various components of a complex process like teaching, one of these being the role of the teacher's knowledge of his/her teaching. He emphasizes two key aspects of such knowledge: that of the subject taught and pedagogical knowledge. Hill, Ball and Schilling (2008) refer to Mathematical Knowledge for Teaching (MKT) as that knowledge used by math teachers in

Statistics Education Research Journal, 13(2), 202-217, http://iase-web.org/Publications.php?p=SERJ

their classrooms, and Schoenfeld and Kilpatrick (2008) have introduced the concept of Proficiency in Teaching Mathematics (PTM) to refer to the professional competence in that specific area.

The models just mentioned characterize teacher knowledge and highlight its importance within the educational field, and the statistics teacher is by no means an exception in this regard. For example, Eichler (2008) conducted a qualitative study with four teachers and found that although courses did not differ in terms of goals and content, the programs actually implemented were significantly different in terms of objectives. According to that author, this difference could be explained by the teacher's conceptions regarding the course and its objectives. This is perhaps why the topic of Evaluation in Statistics Education (EE) has received increased attention in recent years, particularly from the perspective of the answers given by students in evaluations. In relation to this, the Structure of the Observed Learning Outcome (SOLO) Taxonomy proposed by Biggs and Collis (1982) is being used widely to analyze the students' cognitive performance and such taxonomy is also being applied to different content areas. In addition, statistics educators have recently proposed taxonomies to characterize student learning in specific statistical topics or domains (see Aoyama, 2007; Inzunza & Jiménez, 2013; Jones, Langrall, Thornton & Mogill, 1999; Landín & Sánchez, 2010; Reading & Reid, 2007; Vallecillos & Moreno, 2006; Watson & Moritz, 2000).

As can be expected, student answers depend on the type of questions made by teachers. By examining such questions, a good deal of relevant information as to preferred topics and issues, contents, procedures and ideas from a particular course or group of courses can be gathered. Further, statistics exams may address a sample of the cognitive skills that teachers want to tap in their students. For this reason, this study has examined questions that make up a set of written tests in statistics courses, specifically in the career of Education. The study has as its main goal the identification of trends in the types of questions made, and in the content and cognitive skills that are being addressed in such courses.

2. STATISTICS AND ITS EVALUATION IN INTRODUCTORY COURSES

The teaching of statistics courses has evolved through decades. At the beginning, it would focus mainly on mathematics with an emphasis on demonstrations and techniques based on mathematics, as statistics was considered one of the branches of that field. The use of technologies that allowed machine calculations slowly led to changes as calculators were incorporated. This contributed to a move away from mathematics but, while recognizing the importance of calculations, a higher importance was awarded to the interpretation of results, which gave birth to statistics. Then, personal computers and specialized statistical software led to a putting aside of demonstrations and calculations by designers of statistics courses for non-specialists. This implied the understanding that statistics was inseparable from its applications, which in turn led to the major role of such applications in courses for non-specialists.

Many studies (e.g., Cobb, 1992; Moore, 1997) have provided valuable information contributing to transformations in the design of college introductory statistics courses. Such studies have suggested changes in the use of technology, content, teaching, and so forth. A milestone study in this regard has been the Guidelines for Assessment and Instruction in Statistics Education (GAISE) of the American Statistical Association (ASA). According to these guidelines, an introductory statistics course will ideally result in 'statistically educated' students, which means that students must develop both statistical literacy and the ability to think statistically. In order to attain such goals, GAISE (2010) lists six recommendations for the teaching and learning of statistics: (1) to emphasize statistical literacy and the development of statistical thinking, (2) to use actual data, (3) to emphasize conceptual understanding rather than mere knowledge of procedures, (4) to promote active learning in classrooms, (5) to use technology for the conceptual apprehension and analyzing of data, and (6) to use assessments both to improve and evaluate student progress.

The guidelines above seem very important if the objective is to attain a better-prepared student statistics-wise but the kind of evaluation used is also crucial. Instructors can make changes in their teaching methods, may use technology in their classrooms, add modifications

in course content and also use new textbooks, but if they continue to use the same traditional tests (where students are often required to recall or recognize definitions, make calculations and execute algorithms) it is likely that teachers will not be able to know whether the changes introduced have been beneficial or not.

Obviously, the solution will not be to do away with exams but to use them differently, varying both item format and content in order to have a better picture of student achievement, ideally one that is closer to a level of Statistics Literacy and Statistical Reasoning and Thinking. For example, in the Question Bank of the Assessment Resource Tools for Improving Statistical Thinking (ARTIST, https://apps3.cehd.umn.edu/artist/), a few interesting examples of questions that can be used to assess different cognitive levels in statistics can be found. Garfield, delMas and Zieffler (2010) have stated that it is advisable to design tests combining both assessment of understanding of content and also of the cognitive demands implied in the evaluation. In this way, when designing a test, questions will be adjusted to both factors in order to avoid biases either in content or cognitive skill.

On the other hand, Davies and Marriott (2010) recommended that, when designing tests, actual problems should be posed, and some computer output with multiple possible results be presented for students to use and then write a short report. Problems should be closely related to the students' area of specialization so they can make better use of context. It is also advisable that, in order to obtain the best evidence of student achievement, such problems or situations will have enough "free space" for both good students and more disadvantaged ones. Davies and Marriott (2010) also recommend the use of portfolios, to emphasize practical work in the classrooms; to do research that implies discussions, and so forth, as means of diversifying assessments.

The GAISE report highlights the importance of using evaluations not only to assign a grade but also to guide the student's learning. This requires that the student receives feedback, as detailed as possible, on his/her performance on the activity. Such feedback may work as a model for students on how to think and reason about statistics. Clearly, this might be a difficult task to accomplish with large courses, but it is worth the effort if we want to have statistically educated students.

3. THE INVESTIGATION

The study conducted is field research that is exploratory in nature. Teachers of statistics courses at the School of Education of the Universidad Central de Venezuela were invited to submit tests they had used in their courses during the last school year or the last two semesters. Ten teachers complied and handed in 97 different tests, for a total of 978 questions or test items. Teachers who participated had a teaching experience ranging between 4 and 23 years in same or similar courses.

The evaluations belong to a group of courses, namely, Statistics Applied to Education, a year-long program, and also include Mathematics and Statistics Level I, Statistics Level II, and Statistics Level III, taught in a distance program on a semester basis. The annual course is equivalent to the three semester ones, where the first two levels deal with aspects of descriptive and inferential statistics. The level III course includes nonparametric statistics and one-way analysis of variance.

The questions or items in the tests were classified in terms of their assigned weight, the area of statistics being addressed, the context of the question, the type of question or activity and finally the level of difficulty, all of them estimated via expert ratings. The weight or value of each question is normally assigned by the teacher when designing the test. This value is usually printed just beside the question or item on the test sheets. Tests have a maximum possible score of 20 points and the areas or domains evaluated include descriptive statistics, probability and inferential statistics. The contexts in the questions include those related to education, others unrelated to education, and no context (context-free questions). In terms of the type of question or item, these are classified as essay-like, simple selection (multiple choice), true or false and short answers (fill-ins). To estimate the level of difficulty of questions, both the investigator and the raters solved them and then weighted them accordingly. This weight is an estimation

based on the percentage of correct answers that could be expected in a specific test. These categories, in terms of difficulty are: very easy (81-100%), easy (61-80%), average (41-60%), difficult (21-40%) and very difficult (1-20%).

To classify the types of questions according to the cognitive skills to be addressed, an adaptation of the SOLO taxonomy in addition to definitions of literacy, reasoning and statistical thinking by delMas, Garfield, Ooms and Chance (2007) was used. While the SOLO taxonomy has been designed to evaluate student cognitive performance levels, it can also be used in setting course objectives, including required levels of learning (Biggs, 1999). In relation to this, the SOLO taxonomy was adapted (Biggs & Collins, 1982) in order to know the performance levels expected from students. The definitions used were:

- *Unistructural level*. Questions that contain explicit information that when extracted directly from the statement will lead to answers. Comprehension of information in the question may provide a hint to readers as to the procedure to find the answer.
- *Multistructural level.* Questions that require the use of two or more sources of information directly obtained from the statement to arrive at an answer. Data is processed in a sequential way to come up with new information.
- *Relational level*. Questions that require an analysis of the information, establishing relationships between different aspects to integrate the information and generate implications, consequences, predictions or conclusions based on elements from the context. The means to solve questions are not obvious.
- *Extended abstract level*. Questions that demand the use of a general and abstract principle that can be inferred from the information in the statement. Integrating acquired information to generate new information may lead to solving the question.

The SOLO taxonomy enables classification of questions in a general way. The definitions of literacy, reasoning and statistical thinking allow their classification in a more specific area of Statistics Education. To attain this, the definitions of delMas et al. (2007) were used:

- *Statistical literacy.* Question addresses student knowledge of the meaning of statistical terms, use of basic statistical language and instruments, comprehension and use of symbols, recognition and interpretation of data representation (graphs and tables). It has to do with why the data is necessary and how it can be obtained. The question addresses comprehension of basic notions of probability and of how conclusions can be obtained from statistical inferences. It is directed to the students' capability of interpreting and critically evaluating statistical information as well as communicating statistical ideas.
- *Statistical reasoning.* Question seeks to investigate whether the student comprehends statistical ideas, explains statistical processes and interprets statistical results. It involves a relationship between two or more statistical concepts and giving meaning to statistical information. It has to do with making interpretations based on sets of data and with representation of data or making summaries of data.
- *Statistical thinking*. Question addresses knowledge of the process of statistical investigation, the understanding of how models are used to simulate random phenomena and how they can be useful to estimate probability. It implies an adequate comprehension of sampling, of how to make inferences based on samples and how to design experiments with the objective of establishing causality. Also the ability to understand phenomena in terms of their context as well as the ability to plan, evaluate, investigate and arrive at conclusions. It seeks understanding of how, when and why inference tools can be used to support investigations. It also requires the criticizing and evaluating of results of investigations and implies the comprehension of why and how these investigations have been conducted.

All 978 questions were solved by the researcher in order become familiar with the cognitive processes involved in the tasks and also to classify them. To confirm the validity of this classification made by the researcher, the assistance of two professors-raters was requested. These two experts were provided with the set of questions, answers and definitions used for the classification. The experts individually made their own classification in terms of cognitive skills, statistical area or domain addressed, context of questions, question type and level of

difficulty. Initial agreement between the classifications made by the researcher and those by the two expert-raters was approximately 38 percent of the questions. To discuss disagreements between classifications, two meetings with raters were held, and a single classification for all questions was created.

4. RESULTS

4.1. GENERAL CHARACTERISTICS OF QUESTIONS

Most questions (76.8%) included in the study were essay-like, which is in agreement with a long tradition in the area of quantitative testing (e.g., in mathematics, statistics, and so on) at the different levels of education where the expectation is that students "provide" an answer. It is important to say that this type of question is usually not used to address mere knowledge of information or memorization of facts but to evaluate more complex learning in which students are expected not only to provide evidence of what they know but also to show their ability for argumentation and justification of answers.

The remaining questions (23.2%) belong to the type of 'objective' or 'short answer' questions, namely: simple selection (8.4%), true-false (3.8%), and sentence completion (11.0%). It was found that objective and short-answer questions were used as a complement to essay-like questions. These latter ones are the core of examinations and do have a relatively high weight, while the short-answer ones are awarded a smaller weight. Figure 1 shows the results according to the area of statistics to which questions belong.

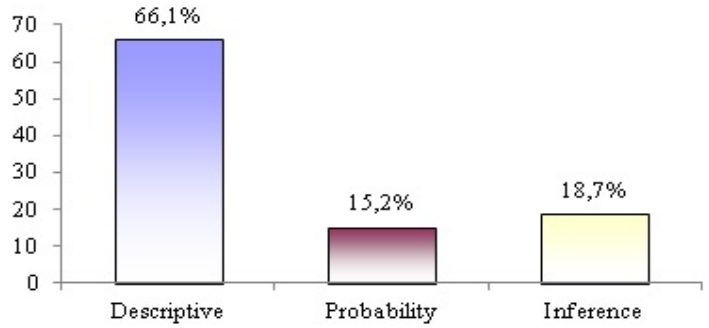


Figure 1. Questions grouped by statistical area addressed.

Two thirds of the questions analyzed belong to the area of descriptive statistics, while the remaining third are divided almost equally between probability and statistical inference. The fact that most questions belong to descriptive statistics may be related to the so-called "school pyramid", which means that such domain of statistics is normally placed at the beginning of university careers. Freshmen courses usually have a higher number of subjects while in the more advanced courses the topics of probability and statistical inference are normally included. This finding could also be due to the fact that programs tend to put more emphasis on descriptive statistics, which covers about half the program contents in most careers. In any case, a significant share of descriptive statistics questions seems to conform to program or student needs and expectations and not to particular teacher preferences.

As for question context, 56% of questions relate to the field of education (figure 2), something which can be expected in a statistics course in this career. Context means the application of statistical knowledge to specific fields, so the fact that work with these courses includes a good share of contexts related to education is expected in order to try to facilitate learning and also to promote possible application of statistical knowledge in the students' future professional life. Needless to say, context is essential for students to make sense of results, and it is especially important in beginners' courses.

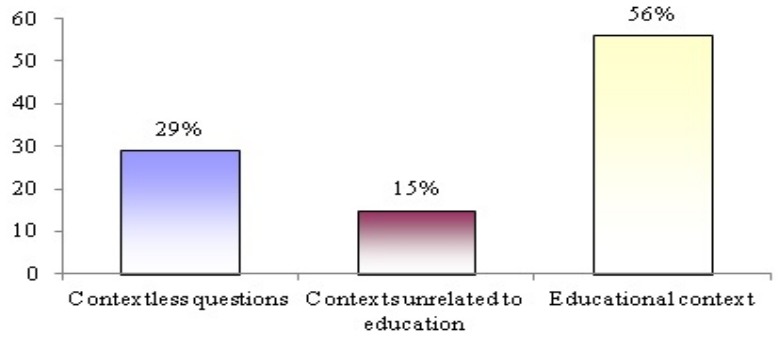


Figure 2. Questions grouped by context.

Preference in terms of educational context is given to achievement. These are questions that often deal with evaluation (grades, test results, assessments) and its analysis via statistical techniques. Other contexts connected with education are student traits, admission, and educational research. However, none of these three contexts gets more than 5% of questions. The meager presence of contexts related to educational research is worth noticing, if we accept the assumption that professional education should be closely related to research issues. Some examples of questions (translated from Spanish) from contexts associated with education are:

Achievement:

Question 51. *It is known that 30% of statistics students usually study in groups. If we want to conduct a study on all statistics students, we set a maximum error acceptability of 2.5% and a 90% confidence. How large should the sample be? (3 points)*

Question 104. *A test has ten simple-selection questions, each one with six options with only one correct. If a student answers the questions randomly, what is the probability for the student to answer at least 1 question correctly? (2 points)*

Student traits:

Question 532. *If you are interested in applying the correlation coefficient (Phi) to know the relationship between the variables 'gender' and 'place of origin' of students, what should you do with these variables in order to make such calculations? (2 points)*

Question 960. *To estimate the average age of the 354 Bolivar Regional Center students, a random sample of 65 people is taken obtaining an approximately normal distribution, with a median of 32 and a standard deviation of 7. Please calculate the confidence interval to estimate the average age in the whole population with a confidence level of 95%. (3 points)*

Fifteen percent of questions belong to contexts unrelated to education (figure 2). This may be a way to give students the opportunity to apply statistical knowledge to contexts other than their field of study. In this case, the context had to do with general aspects of reality and objects, frequently followed by 'time'. However, the percentage of both contexts is quite small. In the first case, it is barely over 5% while the second one is barely over 3%. The remaining ones are all under 2%. Examples of questions in contexts related to education are:

Question 201. *Please say which of the following cases is the frequency and which is the variable: 35 years is the age of 10 of the number of participants.*

Question 854. *A group of students selected 50 animals for an experiment. Students are expected to give them a certain amount of food for a period of 2 weeks. The average weight increase of the animals is 42 grams, with a dispersion of 5 g. Please find a 95% confidence interval to estimate the average increase expected in the population. Please interpret. (3 points)*

It is worth noticing that 29% of the questions were formulated without a specific context in mind, even though it is known that context is crucial for the interpretation and analysis of results. Examples of these questions are the following:

Question 3. *How would you interpret a kurtosis coefficient of -0.216 points?*

Question 30. *In a nominal scale variable, what is the most appropriate measure of central tendency?*

Question 132. *If we think of an interval of 95% confidence this indicates that …:*

Question 187. *What kind of measures can be used to study the heterogeneity of a data set?*

Most of these context-free questions are of the short-answer type, intended to examine theoretical aspects of statistics and devised primarily to elicit information or ask students to perform simple calculations. Certainly, almost all of them could be modified and placed in a specific context as a way of improving question quality and providing a more evocative level of information. Figure 3 below shows that 50% of the questions are given a weight of a point or less for tests with a maximum score of 20 points. This relationship suggests that a good number of questions are assigned a low weight and consequently it could be assumed that there are not many high-demand questions in terms of cognitive skills. It is also possible that questions with a weight of less than one point include multi-part items or are short-answer questions.

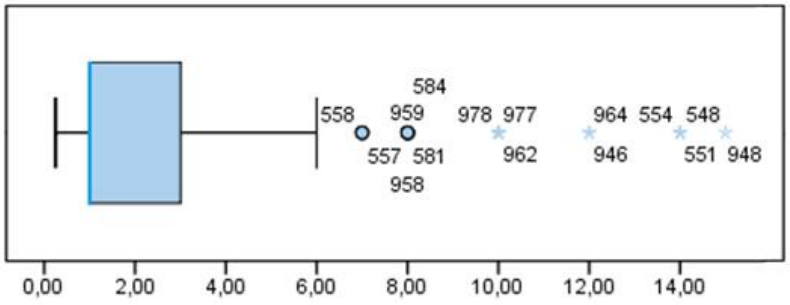


Figure 3. Weight of questions.

Seventy five percent of questions have a weight of three or fewer points, which suggests that high-weight questions are quite uncommon. Obviously enough, if most questions are assigned a low weight, it means that their answers will not require much elaboration and they have a low level of difficulty. In Figure 3, some atypical values are shown. These are questions with weights that range from 10 to 15 points, which equal 50 or 75% of the total possible test score. The fact that these questions are depicted as atypical values means that they are infrequent in the tests analyzed and their weights are considerably different from most of the distribution. In Figure 3, four circles have been marked and eight numbers are shown, which indicates the presence of eight questions considered atypical because of their weights. According to such values, these questions might be considered very difficult, possibly requiring a high level of reasoning or superior cognitive skills on the part of the student. Some examples of this kind of question are the following:

Question 946. *These are the scores from a group of applicants to the Faculty of Pharmacy of the UCV in the areas of Numerical and Verbal Ability:*

Numerical Ability Test	***Verbal Ability Test***
72	*82*
83	*99*
83	*81*
90	*91*
92	*82*

Determine and interpret the relationship between both variables by using two different indexes (12 points).

Question 948. *During the last 10 years the number of teachers pursuing graduate studies at the Higher Education Institute in a municipality is:*

Years	*Number of teachers*
2000	*309*
2001	*280*
2002	*271*
2003	*328*
2004	*310*
2005	*383*
2006	*365*
2007	*482*
2008	*567*
2009	*693*

Based on this information please do the following: Determine the linear regression equation adjusted to the data provided (15 points).

The test items just shown do not seem to demand particularly high levels of reasoning. These are examples of items that require a good amount of calculations in order to arrive at the correct answers and that demand the use of algorithms. It is important to notice that students are allowed to use calculators when writing these tests and such devices feature various routines to complete these statistical tasks. It would seem, then, that the weights of questions have more to do with the number of steps necessary for solutions or with the skills in using the calculator and not with their cognitive demands.

Table 1. Frequency and percentage of questions by level of difficulty.

Level of difficulty	**Frequency**	**Percentage**
Very Easy	363	37.1
Easy	319	32.6
Average	269	27.5
Difficult	23	2.4
Very Difficult	4	0.4
Total	**978**	**100**

According to table 1, less than 3% of the questions can be classified as 'difficult' or 'very difficult'. Therefore, almost all questions are either 'average' or 'less than average', the majority being 'less than average'. Approximately 70% of the questions have been rated as 'easy' and 'very easy', which means, according to the expert raters' criteria, that it is expected that between 61 and 100 percent of the students will correctly solve 70% of the questions in the tests. These results seem consistent with findings in the weights assigned to questions, according to which 75% of questions are assigned three points or less. If a question is assigned a weight of three or fewer points from a total possible score of 20, an assumption can be made that such questions do not demand high cognitive skills for their solutions. It would seem, then, that most questions have a low level of difficulty.

4.2. COGNITIVE ABILITIES ADDRESSED

Definitions from delMas et al. (2007) were used to classify questions according to their cognitive learning potential. The following table shows the results of the classification obtained after discussion with experts:

Table 2. Frequency and percentage according to expected learning

Expected learning	**Frequency**	**Percentage**
Statistical literacy	814	83.2
Statistical reasoning	148	15.1
Statistical thinking	16	1.6
Totals	**978**	**100.0**

According to the classification, 83.2% of the questions evaluate Statistical Literacy, while 15.1% evaluate Statistical Reasoning and 1.6% Statistical Thinking. Most questions are placed at the lowest level of the taxonomy, which means that they are designed to know whether the student comprehends and uses both statistical language and instruments adequately. These results seem to be in agreement with international recommendations according to which it is desirable that statistics courses emphasize *statistical literacy.* However, other levels of the taxonomy appear to be underestimated, at least in the tests analyzed. In this regard, Salcedo (2012) has shown that most program objectives have to do with *statistical literacy*, so a valid expectation would be that tests will tend to match instructional objectives. To illustrate the type of questions in each level of the taxonomy, here are a few examples:

Statistical literacy

Question 4. *The following are the scores from Section 12 students in their first mid-semester test of Statistics III in the 2004-II semester.* ***Scores***	*4-6*	*7-9*	*10-12*	*13-15*	*16-18*
Students	*3*	*8*	*9*	*15*	*12*

Please calculate and interpret the most frequent score obtained by the class. (2 points)

Question 11. *The branch of statistics that describes only a group's characteristics without making inferences or predictions about the population is called: _____ (0.5 points)*

Question 95. *Two events, A and B, are said to be incompatible when:*

a) The probability of A conditioned by B is equal to the probability of A.
b) The probability of the intersection is equal to the product of the probabilities.
c) A and B can not happen simultaneously.
d) The combination of A and B is the actual event.

Question 232. *The following are the scores from a group of 24 students of the School of Education in the final exams of Statistics I and II during the 2003-I semester.*

STATISTICS I											
	2	*6*	*0*	*0*	*9*	*0*	*5*	*0*	*1*	*9*	*2*
STATISTICS II											
	1	*6*	*3*	*2*	*8*	*1*	*3*	*0*	*5*	*0*	*4*

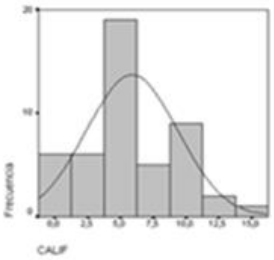
Based on this data: Draw the dispersion diagram and analyze the kind of relationship there is. (3 points).

The questions above are similar in the sense that in order to solve them some basic information and statistical knowledge will be required. But in a few cases it may be necessary to make more than one calculation.

Statistical reasoning

Question 76. *The hypothesis that sex and achievement are unrelated will be examined in this course. For this, tests will be randomly sampled and analyzed from professor Sanadie's courses of the last five years. For females, it is known that 10 have passed and 40 have failed, while 20 males have failed and 30 have passed. Using a level of significance of 5%, establish fitting conclusions.*

Question 307. *Test results from students of EUS Statistics I course from the Barquisimeto Center in their first mid-term evaluation are shown below. Describe the group's performance based on each central tendency measure and then calculate their position in terms of score distribution. Determine the agreement between statistical results and the graph below.*



N		48
Mean		5,88
Median		6,00
Mode		6
Percentile	10	1,00
	25	3,25
	75	8,75
	90	10,00

Question 340. *The Academic Aptitude Test results from 36 senior San Jose High School students, classes A and B are shown below:*

	A	B
N		
Mean		
Median	60,00	61,50
Mode		
Standard deviation		
Variance	96,879	103,820
Skewness (α^3)	0,386	
Kurtosis (α^4)		−1,650
Min value	44	46
Max value	75	73

A	B
65	70
73	62
52	50
48	72
74	66
61	66
60	51
75	48
47	45
73	50
48	44
53	73
46	61
60	66
56	46
60	71
50	66
60	55

Please draw conclusions for both classes based upon students' performance in such test.

As can perhaps be easily noticed, these questions demand more than the remembering of some terminology or the making of calculations. Students are expected to do thinking that involves mastery of statistical ideas, and to relate and interpret them. The chief operations are not calculations, which could be rather simple, but to relate two or more concepts and thus generate statistical information.

Statistical thinking

Question 546. *Based on the data below, think of an educational study where the population, the sample, the statistics and the parameters are clearly established so the following procedure can be conducted: Confidence interval for the population's average (a table with information on the final scores of 20 Statistics III students from 4 regional EUS centers is provided).*

Question 576. *Put forward a research situation in which the Population Proportion Hypothesis Test may be used (5 points).*

Question 582. *Put forward a research situation in which the Population Average Difference Hypothesis Test may be used (5 points).*

The above are some questions classified as statistical thinking questions. These examples are typical; some other similar ones tend to have the same structure but either the data or the parameters have been modified. The three examples above are similar in the sense that they require the student to present a "research situation" or an "educational study". If this activity is carried out, it would be necessary for the student to know about the use of statistics in research processes. They must also know how, when and why to use the inference tools. In this case, the student might generate an exercise-type situation similar to those in the statistics manuals in order to illustrate the use of such instruments. However, the question implies the posing of a research problem and the planning of an investigation that includes the use of statistics.

Now, after classifying the questions based on the SOLO taxonomy in order to estimate the kind of response expected by teacher, 66% was placed at the unistructural level while 27.9% was considered multistructual, 5.7% was placed at the relational level and 0.4% at the extended abstraction level (see Table 3).

Table 3. Frequency and percentage of questions according to the SOLO taxonomy.

Level	Frequency	Percentage
Unistructural	645	66.0
Multistructural	273	27.9
Relational	56	5.7
Extended Abstraction	4	0.4
Totals	**978**	**100**

The results in Table 3 show that the majority of questions analyzed demand lower-level cognitive skills. Students may give answers based mainly on the information explicitly found in textbooks and most answers can be obtained by just following a procedure or algorithm. In addition, 27.9% of questions can be answered by applying two or more established procedures to the information provided in the exam. It would seem, then, that such questions are useful if one wants to know whether the students can remember or reproduce information given in class or show ability to describe, classify or comprehend facts without much systematization. According to Salcedo (2012), most program objectives in these courses have to do with information provided in class (Unistructural Level) and with the unsystematized comprehension of facts (Multistructural Level) from the SOLO taxonomy. Consequently, the tests seem to conform to program objectives. The following are examples of questions classified at all levels of the taxonomy.

Unistructural level

Question 1. *Which scales refer to the qualitative level of measurement?*

Question 55. *Some characteristics from a binomial experiment are: (Choose the right answer)*

a) Symmetrical distribution; $p \neq q$ and two posible results in each trial.
b) Discrete variable; symmetrical distribution and mathematical expected value equal to 1.
c) Continuous variable; $p = q = 0.5$ and $p + q = 1$
d) $p + q = 1$; independent trials and asymptotic.
e) N identical trials; each trial has two possible results, where p and q remain constant.

Question 243. *A Dean wishes to conduct a survey to determine the proportion of students in favor of changing the institution's location. Since surveying all 3,540 students in a*

reasonable period of time is almost impossible, please determine the size of sample (number of students to be interviewed) needed to estimate the proportion of students in favor with a maximum acceptable error of 0.05 and a confidence level of 95%. (4 points)

Question 460. *The following are the grades of 20 students from a statistics course: 12, 15, 10, 13, 16, 12, 18, 18, 20, 8, 6, 17, 14, 18, 6, 7, 5, 4, 12. Please make a frequency distribution of 5 classes.*

The four questions above explain the activity to be carried out and the information needed to accomplish it. Because of this, such questions have been classified as belonging to the Unistructural Level. Generally, these are questions that only demand comprehension of information and basic knowledge to arrive at the answers.

Multistructural level

Question 6. *The following are the scores attained by students of Class 12, Statistics III in the first mid-term test, 2004-II semester.*

Grades	*4-6*	*7-9*	*10-12*	*13-15*	*16-18*
Students	*3*	*8*	*9*	*15*	*12*

Please calculate and interpret the coefficient of variation (3 points).

Question 42. *A population has been put together. The data is: 3, 7, 11 and 15. Considering every possible samples of size two without replacement that can be extracted from such population, please determine the populational average.*

Question 195. *In a group of 60 students admitted into UCV's School of Education for the 2003-4 semester, it was noticed that the average academic index was 66.253 points with a standard deviation of 8.435 points; in addition, their average score until grade 11 was 13.57 points with a standard deviation of 2.31 points. In which of the two scores is the data more heterogeneous? (2 points).*

Question 562. *A student answers randomly a true-or-false test made of three questions. Construct the probability distribution for the random variable "Number of questions correctly answered". Calculate and interpret the expected value*

.

In the four questions above it is necessary to make at least two steps or calculations to arrive at the answers, but these steps are expected to be known by the student. For example, in question 6 students are required to calculate and interpret the coefficient of variation. In order for them to arrive at the correct answer they must calculate the arithmetic average, then calculate the typical deviation, and finally find the coefficient of variation. Once this coefficient has been obtained, it must be interpreted according to the homogeneity or heterogeneity of the values conforming to its arithmetic average. All these measures are calculated via algorithms whose formulas are usually available during tests.

Relational level

Question 75. *A professor of Statistics Applied to Education is interested in evaluating the effectiveness of a new teaching strategy. Once the class has taken the first written test, the professor starts to evaluate her 'experiment' (The test I is using the old strategy and test II is using the new one). Then, the professor chooses to select 8 random students and asks them about their scores in the first and second tests, respectively. The information given to her is shown in this table:*

Student	Test I	Test II
A	10	15
B	5	13
C	12	15
D	16	20
E	8	14
F	11	12
G	18	17
H	20	19

Please help the professor evaluate statistically her proposal using a 5% level of signification.

Question 175. *The Julio Calcaño High School guarantees that their 11th grade students get an average score above 56 points in their Verbal Ability Academic Aptitude Test. A random study using the last 4 years and 52 students reveals an average of 49 with a 7.8 point standard deviation. Does this study support what the school 'guarantees'? Make conclusions based on a 5% level of signification. (3 points)*

Question 538. *In a sample of 150 elementary education youngsters from public schools in the municipality of Guanta, it has been observed that 87 students have difficulty with basic mathematical thinking. Experts from the Ministry of Education have always declared that the percentage of kids with this problem at the national level is less than 55%. Does the information from the Guanta student sample support the experts' opinion at the 1% level? (3 points)*

Question 862. *The average score in the course World History at the "País Portátil" High School is 11.5 out of 20 with a variance of 7.25 points^2. This year the teacher who taught the course for over 25 years has just retired and a new instructor took over. This new teacher has used new strategies to teach the course. In order to evaluate the teacher's job the School's Headmaster selected randomly a group of 40 students whose average score was 13.1 points. How would you assess the performance of the new teacher? Please use $\alpha = 5\%$.*

Even though in many of the relational level questions exemplified above it is possible that students may easily identify the procedure to arrive at answers, the questions demand the generation of implications or conclusions. For this, they must establish relationships between various elements in the text. After arriving at an answer, students must analyze the information; make connections and conclusions, taking into account the context. For example, in Question 862 students are even required to make a value judgement since they are asked about their opinion concerning a new professor's performance, always based upon the statistical evidence available. It is worth mentioning that almost every question at this level refers to content of statistical inference.

Extended abstract level

Question 579. *Please put forward a research situation in which the Population Average Hypothetical Test may be used (μ). (5 points)*

Question 613. *Put forward a research situation in which the Population Proportion Difference Hypothesis may be used (P_1- P_2). (5 points)*

Just as with the case of statistical thinking, questions placed at the Extented Abstract Level are uncommon. The examples shown above refer to situations where students are asked to formulate a 'real' research situation, since for them it is important to apply knowledge and think about an authentic research problem that demands the use of specific techniques of inferential statistics.

In summary, results show that tests are mostly comprised of questions of statistical literacy whose solutions can be found by comprehending information in the items and by following a procedure or algorithm studied in class, as can be seen in Figure 4.

Further, most questions are considered to belong to the domain of statistical literacy both at the Unistructural and Multistructural levels. Therefore, it can be said that such questions basically address skills related to the reproduction of knowledge. The main characteristics of the questions analyzed are the following: they address aspects of descriptive statistics, are essay-like (students must write their own answer), the preferred context is the educational one, mainly *student achievement* (tests, gradings). Most questions have a maximum weight of three points out of 20, and most questions are rated as *very easy* or *easy,* and consequently they are expected to be solved correctly by at least 61% of students.

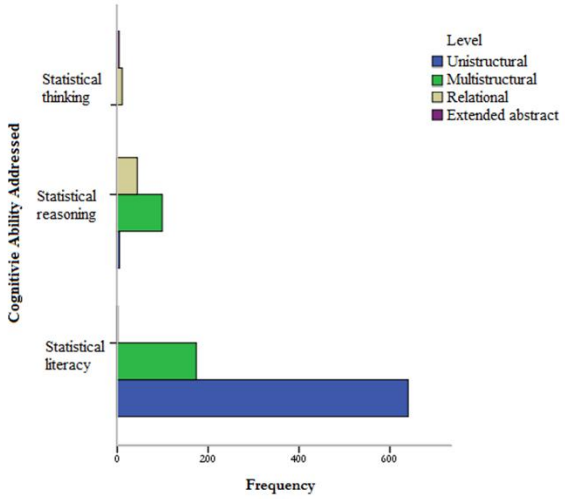


Figure 4. Questions classified in both taxonomies.

The majority of questions have been placed at the Unistructural and Multistructural levels in terms of expected learning. Consequently, it could be said that these questions basically address abilities related to knowledge reproduction and which try build on very basic knowledge of statistics. They usually have to do with calculations that can be made by following one or two procedures. Questions that address more complex statistical ideas or the application of statistical knowledge to educational research are scarce.

5. REFLECTIONS

Based on the premise that a test is a means of gaining information about how much learning and what kind of it our students have acquired, results of this study show that the tests analyzed have been devised in order to know whether the students have comprehended some basic statistical concepts and if they are capable of remembering content studied in class, which basically refers to the standard calculations. As Brown and Glasner (2003), Watts, García-Carbonell and Llorens Molina (2006) have stated, practices and criteria concerning evaluation tend to orientate students in terms of where they are expected to direct their learning activities. Specifically, in statistics for education courses, students must direct their efforts towards the mastery of statistics terminology and the application of basic statistical instruments. They must also address the representation of data and making of calculations via algorithms of Descriptive Statistics. The strong bias towards topics and issues of statistical literacy is partially in agreement with GAISE recommendations (ASA, 2010) as regards emphasis on that domain in courses for non-specialists. The other domains of statistical reasoning and thinking are often left unattended although it is known that without some emphasis on these latter domains it might be impossible to comply with the goals and demands set in the second part of these recomendations, more related to the development of statistical thinking.

The fact that most questions are placed at the lower levels of the taxonomies suggests that the core of the expected learning processes focuses mainly on content and procedures, not on

fundamental statistical ideas, their meaning and application in various contexts. Both taxonomies point to the need for a clearer definition of the level of comprehension that students are expected to achieve, of aligning evaluation with the learning objectives and with the teaching approaches or methods.

An important issue arising from this study is that test questions seem to be in agreement with the demands of course programs. Due to this, it would be necessary to revise course programs in order to introduce changes and modifications based on international recommendations. It would also seem advisable to conduct research on activities usually carried out in classrooms, on homework and other evaluation-related activities, in terms of cognitive level demands. Since we have dealt with an introductory statistics course, it is only reasonable that emphasis is placed on issues of statistical literacy, but other deeper levels of comprehension should not be neglected. The main goal should be the development of statistical thinking and to attain this, students should engage in learning situations in which they are able to come closer to fundamental statistical ideas and go beyond definitions, formulas and calculations.

ACKNOWLEDGMENTS

Our thanks to colleagues Amalio Sarco Lira, Jesús González, and Santiago Inzunza for their suggestions and comments on an earlier version of this article.

REFERENCES

American Statistical Association (2010). *Guidelines for Assessment and Instruction in Statistics Education – GAISE. College Report*.
[Online: http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf]

Aoyama, K. (2007). Investigating a hierarchy of students' interpretations of graphs. *International Electronic Journal of Mathematics Education*, *4*(3), 298–318.
[Online: http://www.iejme.com/032007/d10.pdf]

Biggs, J. B. (1999). *Teaching for quality learning at university*. Buckingham. Open University Press.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Brown, S., & Glasner, A. (Eds.). (2003). *Evaluar en la universidad. Problemas y nuevos enfoques*. Madrid: Narcea.

Cobb, G. (1992). Teaching statistics. In L. A. Steen (Ed.), *Heeding the call for change: Suggestions for currricular action*, *MAA Notes, Vol. 22* (pp. 3–33). Washington, DC: Mathematical Association of America.

Davies, N., & Marriott, J. (2010). Assessment and feedback in statistics. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 3–19). Chichester, UK: Wiley.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58.
[Online: http://iase-web.org/documents/SERJ/SERJ6(2)_delMas.pdf]

Eichler, A. (2008). Teachers' Classroom Practice and Students' Learning. In C. Batanero, G. Burrill, C. Reading, and A. Rossman (Eds.), *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and IASE 2008 Round Table Conference.* Monterrey, Mexico, International Commission on Mathematical Instruction and International Association for Statistical Education.

Garfield, J., delMas R., & Zieffler, A. (2010). Assessing important learning outcomes in introductory tertiary statistics courses. In P. Bidgood, N. Hunt& F. Jolliffe, (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 75–86). Chichester, UK: Wiley.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*(4), 372–400.

Inzunza, S., & Jimenez, J. V. (2013). Caracterización del razonamiento estadístico de estudiantes universitarios acerca de las pruebas de hipotesis. *Revista Latinoamericana de Investigación en Matemática Educativa*, *16*(2), 179–211.

Jones, G. A., Langrall, C. W., Thornton, C. A., & Mogill, A. T. (1999). Students' probabilistic thinking in instruction. *Journal for Research in Mathematics Education, 30*(5), 487–519.

Landín, P. R., & Sánchez, E. (2010). Niveles de razonamiento probabilístico de estudiantes de bachillerato frente a tareas de distribución binomial. *Educação Matemática Pesquisa*, *12*(3). [Online: http://revistas.pucsp.br/index.php/emp/article/viewArticle/4842]

Moore, D. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*(2), 123–137.

Reading, C., & Reid, J. (2007). Reasoning about variation: Student voice. *International Electronic Journal of Mathematics Education*, *2*(3), 110–127. [Online: http://www.iejme.com/032007/d1.pdf]

Salcedo, A. (2012). *El examen de Estadística: Tendencias y reflexiones*. Trabajo de ascenso no publicado. Caracas: Universidad Central de Venezuela.

Schoenfeld, A. H., & Kilpatrick, J. (2008). Towards a theory of profiency in teaching mathematics. In D. Tirosh & T. Wood (Eds.), *Tools and processes in mathematics teacher education* (pp. 321–354). Rotterdam: Sense Publishers.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Vallecillos, A., & Moreno, A. (2006). Estudio teórico y experimental sobre el aprendizaje de conceptos y procedimientos inferenciales en secundaria. *Tarbiya: revista de investigación e innovación educativa*, *38*, 61–78.

Watson, J. M., & Moritz, J. B. (2000). The longitudinal development of understanding of average. *Mathematical Thinking and Learning*, *2*(1&2), 11 – 50.

Watts, F., García-Carbonell, A., & Llorens Molina, J. A. (2006). Introducción. In F. Watts & A. García-Carbonell (Eds.), *La evaluación compartida: Investigación multidisciplinar* (pp. 61–76). Valencia: UPV.

AUDY SALCEDO
School of Education
Universidad Central de Venezuela
Depto. de Estadística e Informática
Edificio de Trasbordo
Ciudad Universitaria
Los Chaguaramos. Caracas, Venezuela
audy.salcedo@ucv.ve