PSYCHOLOGY STUDENTS' UNDERSTANDING OF THE CHI-SQUARED TEST

Gustavo R. CAÑADAS[1], Carmen BATANERO[2],
Carmen DIAZ[3] and Rafael ROA[4]

TITRE

Compréhension du test du chi-carré par des étudiants de psychologie

ABSTRACT

This paper describes a study of the competence to carry out the Chi-squared test, which is frequently used in psychology and education, in a sample of psychology students after studying the subject. Using an open-ended problem that is solved with the help of an Excel program, we analyse the setting of hypotheses, identification of the statistics and *p*-value, decision taken and interpretation of results. Some implications for teaching are also included.

Keywords: *statistical tests, Chi-squared, understanding.*

RÉSUMÉ

Ce document décrit une étude de la compétence pour effectuer le test du chi-carré qui est fréquemment utilisé en psychologie et en sciences de l'éducation. On observe un échantillon d'étudiants en psychologie après qu'ils aient étudié le sujet. Par la considération d'un problème ouvert, qui est résolu avec l'aide d'un programme Excel, nous analysons la formulation des hypothèses, l'identification de la statistique du chi-carré et de la *p*-valeur, la décision prise et l'interprétation des résultats. Quelques implications de cette étude pour l'enseignement sont également abordées.

Mots-clés : *tests statistiques, chi-carré, compréhension.*

1 Introduction

While students may easily learn formulas and apply statistical procedures, they may not attain the knowledge required to adequately solve problems using inferential statistics; consequently it is important to investigate the students' difficulties as well as the factors that support the students' skills to correctly apply statistical methods (Alacaci, 2004). One main inferential tool is the statistical test that aims to state the evidence in a sample against a previously defined null hypothesis. Statistical tests are difficult to understand by students, because performing these tests requires students to understand and be able to relate many abstract concepts such as the concept of sampling distribution, the significance level, the null and alternative hypotheses, and the *p*-value (Castro Sotos *et al.*, 2007).

A common statistical analysis is the Chi-squared test, also referred to as χ^2 test, in which the sampling distribution of the test statistics when the null hypothesis is true is a Chi-squared

[1] Universidad de Granada. España, grcanadas@ugr.es
[2] Universidad de Granada, España, batanero@ugr.es
[3] Universidad de Huelva, España, carmen.diaz@dpsi.uhu.es
[4] Universidad de Granada, España, rroa@ugr.es

distribution. Two cases where this happens are the tests used to examine the homogeneity of several samples and to check the association of two variables from data presented in a contingency table.

Although previous literature has analysed the students' difficulties in relation to hypothesis testing, such studies have relied on multiple-choice items, or only analysed parametric tests for the mean of a normal population. They have also focused on the understanding of the concepts involved in hypothesis testing, rather than on studying the students' competence to complete all the steps in this procedure.

To complement the existing research on this point the objective of our study was to evaluate how psychology students perform the different steps required in a Chi-squared test of homogeneity, after having studied the topic. This test is applied to a single categorical variable from data collected from two or more different populations and is done to determine whether several populations are similar or homogeneous in some characteristics. Our aim is to analyse the different steps needed to perform a complete Chi-squared homogeneity test: setting the hypotheses, identifying the Chi-square and *p*-value with the help of computation software, making a decision, and interpreting the results.

To follow, we first present the bases for this study, then describe method and discuss the results. Some implications for teaching are also included.

2 Theoretical background

2.1 Understanding statistical tests

In the past 50 years hypotheses tests have become predominant in psychological and educational research. In spite of this wide use, the interpretation and application of significance tests have been controversial since their creation, and debates about their use have become popular after Cohen's (1994) paper in the journal *American Psychologist*. Moreover, psychological and educational research has shown widespread misconceptions among both students and scientists who use statistical inference in their work (see Harlow, Mulaik & Steiger, 1997 or Batanero, 2000 for a survey). The prevalence of these errors lead to a paradoxical situation where, on one hand, a significant result is required to get a paper published in many journals and, on the other hand, significant results are misinterpreted in these publications (Falk & Greenbaum, 1995; Lecoutre & Lecoutre, 2001).

The controversy related to the use of statistical tests has been very strong within some professional organizations in the past 15 years (e.g., Ellerton, 1996; Robinson & Levin, 1997; Levin & Robinson, 1999; Wilkinson, 1999; Batanero, 2000; Batanero & Díaz, 2006). Consequently these institutions suggested important changes in their editorial policies regarding the use of statistical significance testing. At the same time the controversy raised the interest of educational researchers who started studying the understanding of statistical tests on students (see Castro Sotos *et al.*, 2007 for a review).

A particularly misunderstood concept is the level of significance, α, which is defined as the probability of rejecting a null hypothesis, given that it is true. The most common misinterpretation of this concept consists of switching the two terms in the conditional probability that serves to define the significance level; that is, the error consists of interpreting the level of significance as the probability that the null hypothesis is true, once the decision to

reject it has been taken. This mistake was described in Birnbaum (1982), who reported that his students found the following definition reasonable: "A level of significance of 5% means that, on average, 5 out of every 100 times we reject the null hypothesis, we will be wrong". Other researchers found the same error. For example, Falk (1986) found that most of her students believed that α was the probability of being wrong when rejecting the null hypothesis. Similar results were found by Vallecillos (1994) in university students and by Haller and Krauss (2002) in university lecturers involved in the teaching of research methods. More specifically they found that 4 out of every 5 methodology instructors had misconceptions about the concept of significance, just like their students.

The level of significance is not the only concept misunderstood in significance testing, but there is also confusion between the roles of the null and alternative hypotheses as well as between the statistical alternative hypothesis and the research hypothesis (Vallecillos, 1994; Chow, 1996). The null hypothesis is assumed to be true, it is planned to be rejected and determines the sampling distribution of the statistics. However, although Vallecillos and Batanero's (1997) participants study agreed on the theoretical idea that we establish a null hypothesis with the intention of finding evidence against it, they were inconsistent when asked to define the null and alternative hypothesis for a specific problem, since they exchanged both hypotheses. Moreover part of the students raised their hypotheses in terms of the statistics (for example, the sample mean) instead of using the parameter (the sample population).

Castro et al. (2007) and Vera, Diaz, and Batanero (2011) informed about the students' difficulties in identifying the population under study. As regards the logic underlying the statistical tests, Vallecillos (1994) reported that many students in her research believed that correctly carrying out a test proved the truth of the null hypothesis, as in the case of a deductive procedure. The same error was also described by Liu and Thompson (2009) when interviewing eight high school statistics teachers, who did not seem to understand the purpose of statistical tests as mechanisms to carry out statistical inferences.

All these difficulties are related to the understanding of the underlying logic of statistical tests (Harradine, Batanero, & Rossman, 2011). These authors suggest that statistical inference consists of three distinct, but interacting, fundamental elements: (a) the reasoning process, (b) the concepts and (c) the associated computations. While the computations are easily carried out today with the help of software, Harradine, Batanero and Rossman claim that the teaching of the logic of statistical tests and underlying concepts is still an open problem.

All the above research focussed on parametric tests; frequently on testing the value for the mean of a normal population and never centred on the understanding of the Chi-squared test of homogeneity, which is the focus of the present study.

2.2 Mathematical objects and semiotic conflicts

Our research was supported by some theoretical ideas from the onto-semiotic approach developed in different works (Godino, 2002; Godino, Batanero & Font, 2007). These authors distinguish the following types of primary mathematical objects, that are organised in either institutional (epistemic) or personal (cognitive) configurations: (a) Problem-situations (in our work, deciding about the homogeneity of several samples); (b) Language (terms, expressions, notations, graphics) used in mathematical work, such as the expressions H_0 and H_1 to describe the hypotheses; (c) Concepts, such as statistics and parameter, level of significance; (d) Propositions, properties or attributes, such as the fact that the null and alternative hypotheses

are complementary; (e) Procedures (operations, algorithms, techniques, such as computing the Chi-squared statistics); and (f) Arguments used to validate and explain the propositions and procedures, including deductive and inductive arguments.

Another component in the model is the idea of semiotic function. Godino (2002) generalizes the notion of representation, by taking the idea of semiotic function from Eco: "*there is a semiotic function when an expression and a content are put in correspondence*" (Eco, 1979, p.83). The authors interpret meaning as the content of any semiotic function, that is to say, the content of the correspondences (relations of dependence or function) between an antecedent (expression, signifier) and a consequent (content, signified or meaning), established by a subject (person or institution) according to distinct criteria or a corresponding code. The content of the semiotic function could be a personal or institutional object; it could be a concept–definition, a problem–situation, a procedure, an argument, or a linguistic element.

To discriminate between institutional and personal meaning for a same mathematical expression, the onto-semiotic approach introduces the idea of semiotic conflict. This idea has been introduced in the onto-semiotic approach as an explanation of students' errors, difficulties and obstacles in the learning of specific mathematical content, and in general, of difficulties arising in classroom communication. A *semiotic conflict* is any disparity or difference of interpretation between the meanings ascribed to an expression by two subjects (people or institutions). In this work the idea of semiotic conflict will be used to explain the students' difficulties in carrying out the Chi-squared test.

3 Method

The sample consisted of 92 psychology students, who followed a course in "Data analysis in psychological research" (first year of psychology degree) during the academic year 2010-2011. Four theoretical and two practical 1 hour long sessions were devoted to the study of contingency tables, including the Chi-squared test of homogeneity. The period of time devoted to this theme was one month, and the evaluation was carried out a week later. These students had also studied the bases of inference (sampling, differences between random and statistical variable, statistics and parameter, sampling distributions, including the normal, t, Chi-squared and F distributions, the confidence intervals, the logic of statistical tests and some parametrical statistical tests (differences of means and proportions) in the first semester of the same year for a period of four months. Moreover they had studied descriptive statistics and a more elementary introduction to sampling an inference in the high school, the previous year (for another semester).

The problem analysed in this report (Figure 1) was solved by the participants as part of their final assessment in the course. Students were given an Excel program where they could get the value for the Chi-square statistics, after entering the table data; the program did not provide the degrees of freedom; although, once students provided this value, the program computed the associated *p*-value.

In the problem, the students had to carry out a test of homogeneity and complete all the different steps. Every hypothesis test requires a null hypothesis and an alternative hypothesis to be stated. These hypotheses should be stated in such a way that they are mutually exclusive. That is, if one is assumed to be true, the other must be rejected as false. Consequently, to solve question (a) the student needs to interpret the statement, and remember that in a

homogeneity test we try to decide whether the data come from different populations or not. A possible correct formulation of the hypotheses is as follows:

H_0: the proportion of people perceiving each word as emotionally positive, negative or neutral is the same.

H_1: some of the proportions of people perceiving each word as emotionally positive, negative or neutral are different.

To answer the second question, the student can use a computer program, which provides the value Chi = 10.81, although he/she must specify the degrees of freedom. In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. The degrees of freedom in the problem is computed by multiplying the number of rows minus one, by the number of columns minus one, that is df = (3-1)x(3-1)=4. Once this value is provided, the program calculates the *p*-value = 0.029.

In question (c) the student has to remember the logic underlying statistical tests and how to build the acceptance and rejection regions, the definition of the *p*-value and the significance level and conclude that the result is not statistically significant for the critical value given in the problem. The last step is to interpret the results: at a 0.01 significance level we cannot accept that the proportion of people perceiving one or more words as emotionally positive, negative or neutral is different (one or more words has different emotional component) and therefore, we must accept that these proportions are identical, so that the data come from identical populations (the three words have the same emotional component).

Throughout an experiment, a psychologist selects three words and decides to assess their emotional component in a random sample of students. He independently presents each word to 100 subjects and records whether the word is perceived as emotionally positive, negative or neutral. In view of the results, should the psychologist consider that all the three words have the same emotional component?

Perception of the emotional component	Word 1	Word 2	Word 3
Positive	26	45	32
Negative	32	27	38
Neutral	42	28	30

a. Establish the adequate hypotheses to perform a Chi-squared test of homogeneity.
b. Compute the test statistics and the associate *p*-value using the software.
c. Decide whether the psychologist should reject or not the null hypothesis at a significance level α=0.01.
d. Interpret the implications of your result in this research.

FIGURE 1 – *Problem proposed to the students*

4 Results

Once the data were collected a qualitative analysis of the responses to each question was carried out, to classify the responses, which are presented below. We denote as CR.n, PR.n and IR.n the correct, partially correct and incorrect responses.

4.1 Establishing the hypotheses

The choice of adequate hypotheses is the first step in carrying out a statistical test and determines the results of the complete process. However, Vallecillos (1994) suggested that defining the null and alternative hypotheses presents great complexity for students, who are unable to identify the most appropriate statement for each case. In Table 1 we present an example for each type of response to question (a). We found the following categories:

CR.1. The student uses an appropriate notation for the null and alternative hypotheses, and makes explicit reference to the parameter "population proportion" in a symbolic way to establish the hypotheses, which are correctly chosen.

CR.2. The student refers to the concept of distribution (instead of referring to the population parameter) and does not use a symbolic expression. The formulation of the null and alternative hypotheses is correct.

TABLE 1 – *Examples for each category of responses in question a*

Category	Example
CR.1	H_0: $p_{1j} = p_{2j} = p_{3j}$ H_1: One or more populations are different
CR.2	H_0: The three populations have identical distribution for variable Y H_1: The distribution for variable Y in some of these populations is different
PR.1	H_0: $\mu_1 = \mu_2 = \mu_3$ H_1: $\mu_n \neq \mu_m$, for some n, m belonging to $\{1,2,3\}$
PR.2	H_0: The three words have the same emotional component H_1: Some words have different emotional component
PR.3	H_0: Emotional component$_1$ = Emotional component$_2$ = Emotional component$_3$ H_1: $EC_1 \neq EC_2 \neq EC_3$
IR.1	H_0: The given words influence the emotions H_1: The given words do not influence emotions
IR.2	H_0: 10,81
IR.3	H_0: $X_1 \neq X_2 \neq X_3$ H_1: $X_1 = X_2 = X_3$
IR.4	H_0: The variables in columns are not associated H_1: The variables in rows are associated
IR.5	H_0: $p_{1j} = p_{2j}$ H_1: different
IR.6	H_0: $p_{1j} = p_{2j} = p_{3j}$ H_1: $p_{1j} \neq p_{2j} = p_{3j}$

PR.1. The null hypothesis (equality of populations parameters) and the alternative hypothesis (difference between some of these parameters) are correctly established, but the student refers to the population mean, instead of using the proportions, in spite of this, there is not much sense in using the mean on qualitative data. Probably the fact that the frequencies in the contingency table are numbers leads the student to forget the difference between quantitative and qualitative data.

PR.2. Language used to state the hypotheses is imprecise, and consequently we cannot be sure that the student distinguishes the variable from its distribution. Although the null hypothesis is correctly formulated, the student does not refer to the concept of population or sampling distribution, which according to Harradine *et al.* (2011) are more abstract than those of population or sample.

PR.3. The student chooses the correct null hypothesis, but the alternative hypothesis is incorrect, since in the homogeneity test it is only required that some of the populations are different (instead of all of them being different). Similarly to Vera *et al.* (2011), the two hypotheses do not cover the parametric space and therefore are not complementary.

IR.1. The student confuses the null and alternative hypotheses (an error described by Vallecillos, 1994 and Chow, 1996). As reported in Vallecillos this confusion turns out to be a serious misconception that obstructs the understanding of the testing process and specially the correct interpretation of its results. Moreover, in this response the student formulates the hypotheses in an independence test. This confusion is reasonable, since the homogeneity Chi-squared test statistics is computed exactly the same as the test for independence, using data from a contingency table. The only difference between the test for independence and the homogeneity test is the stating of the hypotheses since in homogeneity tests a null hypothesis is asserting that various populations are homogeneous or equal with respect to some characteristic of interest against an alternate hypothesis claiming that they are not. A related error was described by Hackett (2010) who reported that some students analyzed nominal data as a Chi-square goodness of fit test when a Chi-square test of association was needed, or vice versa.

IR.2. The student presents the Chi-squared statistics as null hypotheses and gives no alternative hypothesis. When thinking about statistical inference it is necessary for students to clearly differentiate parameter and statistics. The parameter is an unknown value that determines the probability distribution that models the values of a variable in which we are interested (in this problem the proportions of people perceiving each word as emotionally positive, neutral or negative). The statistic is a value computed from the data (in this example the Chi-squared statistics) whose value is known and that is used to test the hypotheses. The hypotheses should be set in terms of the parameters (and not in terms of the statistics); moreover it makes no sense to establish the hypothesis in terms of the statistics, because its value is known. However, students giving this response did not clearly distinguished the concepts of statistics and parameter, and consequently set the hypothesis using the statistics, an error reported by Vallecillos (1994) and Castro Sotos *et al.* (2007).

IR.3. The student exchanges the null and alternative hypotheses, and in the alternative hypothesis, he requires that the three populations have to be different. However, if the study has three or more populations and the test of homogeneity yields a significant p-value we should decide that the populations are not identical, but it is not possible to conclude which of the populations differ without further testing. Moreover, as the null and alternative hypotheses are not complementary, the hypotheses does not cover the parametric space, an error reported by Vera *et al.* (2011).

IR.4. As in response IR.1, the student confuses the homogeneity and independence tests and moreover he/she indicates that the association is produced between the different rows or between the different columns (instead of suggesting that the rows and columns are associated). Our conjecture is that the student might be confusing value (each column or row is a value) and variable, an error that was pointed out by White (1980).

IR.5. The student only uses two populations (instead of three); he may be confusing the number of variables (two variables) with the number of populations (three populations); again we conjecture that this error maybe related to the difficulty in identifying the population under study that was pointed out by White (1980).

IR.6. The null hypothesis is correct; but in the alternative the student only requests the difference between the two first populations and, consequently the hypotheses do not cover the parametric space (Vera *et al.*, 2011).

TABLE 2 – *Results in question a*

Response		Frequency	Percent
Correct	CR.1	29	31.4
	CR.2	14	15.2
Partly correct	PR.1	10	10.9
	PR.2	7	7.6
	PR.3	13	14.2
Incorrect	IR.1	2	2.2
	IR.2	1	1.1
	IR.3	4	4.3
	IR.4	2	2.2
	IR.6	2	2.2
	IR.7	1	1.1
No answer		7	7.6
Total		92	100

In Table 2 we observe that 46.6% of students posed correct hypotheses, 32.7% partially correct hypotheses, and 13.1% incorrect hypotheses. The most frequent answers were correct (CR.1 and CR.2), in which students posed correct hypotheses for the homogeneity test, either in a symbolic or verbal way. The next more frequent responses were partially correct (PR.1), where students set the hypotheses in terms of the population mean, instead of using the population proportion, and (PR.3) in which in the alternative hypothesis all the three populations are assumed to be different (instead of some of them), so that both hypotheses do not cover the parametric space, an error described previously by Vera *et al.* (2011) in a different problem. If we add the responses IR.3 and IR.6 (6 more students), about 20% of the students set hypotheses that were not complementary.

The global results were better than those reported by Vallecillos (1994) and those by Vera *et al.* (2011) with 26% and 33% of participants respectively correctly posing the hypotheses in parametric tests (testing the value for the mean of a normal population).

4.2 Test statistics and *p*-value

In the second step, the students should use the Excel program to get the Chi-squared value, and, after providing the degrees of freedom, interpret the *p*- value that will be returned by the program. Statistical inference involves drawing conclusions that go beyond the data using the empirical evidence that supports these conclusions, which have a degree of uncertainty, quantified in this problem by the *p*-value, which accounts for the variability that is unavoidable when generalising beyond the sample data to the population. In spite of the

apparent simplicity of the task, the students still made some mistakes. Below we describe the different categories of answers (an example for each type of response is presented in Table 3).

TABLE 3 – *Examples for each category of responses in question b*

Category	Example
CR.1	Chi-squared = 10.81; *p*-value = 0.029
IR.1	Chi-squared = 10.81; $p = 0.001$
IR.2	Chi-squared = 10.81; $p = 13.28$
IR.3	Chi-squared = 10.81
IR.4	Chi-squared = 10.81; $p = 0.01$

CR.1. In the correct response, the student used an appropriate notation, and obtained correct values for the Chi-square statistics and for p (after providing a correct value for the degrees of freedom).

IR.1. The student correctly computed the Chi-square value, but confused the degrees of freedom and provided 1 d.f. instead of 4; consequently he/she obtained an incorrect p-value (the p-value corresponding to Chi-squared = 10.81 with 1 d.f.). As early as in 1940, Walker (1940) remarked on the difficulty of the concept of degrees of freedom, which is of great importance to modern statistical theory and which few textbooks had attempted to clarify. Modern statistical analysis makes much use of several very important sampling distributions for which the shape of the curve changes with the number of degrees of freedom, which appears as a parameter. Consequently, if a mistake is made in determining these degrees of freedom from the data, the wrong probability value will be obtained, as in this response. The confusion of the degrees of freedom was also found in the Alvarado's research (2007) when his students worked with the t distribution to compute confidence intervals and in that by Olivo (2008) in students working with the t, Chi-square or F distributions in order to determine the sampling distributions for some statistics.

IR.2. The student correctly computed the Chi-square value and the degrees of freedom. He/she then used the Chi-square table and obtained the critical value 13.28 corresponding to the level of significance α given in the problem statement. Then the student confused this critical value with the p-value without noticing that a probability cannot be greater than one, an error also described by Contreras *et al.* (2010). The student showed a routine learning of the several concepts related to performing a statistical test.

IR.3. The student correctly computed the Chi-square value, but was unable to get the p-value; probably because he did not remember the degrees of freedom as in the response IR.1.

IR.4. The student correctly computed the Chi-square value, and provided the significance level value, instead of the p- value; he probably confuses both concepts, an error described by Vallecillos (1994).

Results are presented in Table 4, where we see that the majority of students correctly obtained the Chi-squared statistics and the associated probability (64.2%) with the help of the statistics program. The most frequent incorrect responses (IR.1) and (IR.3) were related to not remembering the formula for the degrees of freedom, an error described by Alvarado (2007) and Olivo (2008), in the use of the t, Chi-square and F distributions.

TABLE 4 – *Results in question b*

Response		Frequency	Percent
Correct	CR.1	59	64.2
Incorrect	IR.1	14	15.2
	IR.2	6	6.5
	IR.3	12	13.0
	IR.4	1	1.1
Total		92	100

4.3 Making a decision

Making a correct decision involves the understanding of the logic behind statistical tests; this logic is supported by previous understanding of concepts such as distribution, centre, spread, association, uncertainty, randomness and sampling (Harradine *et al.*, 2011).

In question c we only found two types of answers. Correct answers are those in which the student interpreted the results appropriately, and did not reject the null hypothesis with formulations such as "*we cannot reject the populations' homogeneity*". Wrong answers correspond to students who chose to reject the hypothesis, since they made a confusion between the acceptance and rejection criteria. These students showed a poor understanding of the logic behind the statistical tests, in agreement with Harradine *et al.* (2011) who suggested that, while most students may be able to perform the calculations associated with an inferential process, many students hold deep misconceptions that prevent them from making an appropriate interpretation of the result of a statistical test.

TABLE 5 – *Results in question c*

Response	Frequency	Percent
Correct	54	58.7
Incorrect	30	32.6
No answer	8	8.7
Total	92	100

In Table 5 we see that the majority of students (58.7%) made the right decision, with better results than those in Vallecillos's research (1994) although still almost half of students failed to reach a correct decision, an error also reported by Haller & Kraus (2002).

4.4 Interpreting the results

The final step is to interpret the statistical results in the problem context. According to Chaput, Girard and Henry (2011) this is the final step in any modelling process and consists first of translating the mathematical results, then giving them a meaning to create answers to the original problem in the real world, and then again comparing these answers with the model hypotheses. Finally, the answers have to be put into perspective to estimate whether the model was adequate for the real problem. An example for each type of response in this part is presented in Table 6.

TABLE 6 – *Examples for each category of responses in question d*

Category	Example
CR.1	*The emotional component is homogeneous in the different words.*
PR.1	*Looking at the Chi-squared table I got a p-value p=0.029. Consequently, it is not statistically significant since it is higher than 0.01. We accept the homogeneity hypothesis.*
IR.1	*We reject H_0; this means that the three words have different emotional components.*
IR.2	*We should reject the independence hypothesis (since the p-value is close to 0, this indicates a very unlikely Chi-squared value).*
IR.3	*Observing the Chi-squared value we obtain that the probability of getting a value of 13.28 or greater with 4 d.f. is p = 0.029.*

We found the following categories:

CR.1. The student interpreted the results in the context of the problem and made a correct decision. Once the decision of not rejecting the null hypothesis was taken, the student was able to interpret what the implication of this decision in the original problem was (assessing the perception of the words' s otional component in the popul ation of students). Since the null hypothesis was not rejected, all we could say is that we should (provisionally accept) that the emotional component of the three words was identical.

PR.1. The student did not interpret the results according to the problem context, but made a correct decision and a generic (but correct) interpretation of a statistical test. In this response the students completed the work with the mathematical model and showed a good understanding of both the different concepts involved in statistical tests as well as in the reasoning involved in the test, as they took the correct decision. However they were unable to complete the last step in the modelling process and then did not perceived what the implications of the mathematical results in the context of the problem were.

IR.1. The student interpreted the results in the context of the problem, but made an incorrect decision, since they rejected the null hypothesis. These students were able to complete the last step in the modelling process (interpreting the results); however they failed to take the correct decision because of an incorrect understanding of the logical reasoning in a statistical test. Moreover some students giving this response took into account the *p*-value but not α explicitly: "*Since p = 0.029 is very close to 0, the Chi-squared value is very unlikely if H_0 true; we therefore reject the equality of emotional component"*.

IR.2. The student did not interpret the results in the context of the problem, and made an incorrect decision, since they reject the null hypothesis. In this case, both the work with the mathematical model and the interpretation were incorrect.

IR.3. These students neither interpreted the results in the context nor did they make a decision.

In Table 7 we observe that 43.5% of the sample correctly interpreted the results in the context of the problem and then completed all the steps in a modelling process; the remaining students either did not interpret the results or made an incorrect interpretation.

TABLE 7 – *Results in question d*

Response	Frequency	Percent
CR.1	40	43.5
PR.1	5	5.4
IR.1	21	22.8
IR.2	2	2.2
IR.4	1	1.1
No answer	23	25
Total	92	100

5 Discussion and implications for teaching

According to Alacaci (2004), understanding statistics is a universal requirement for students in the social sciences; it is important therefore to develop in these students the appropriate ability to identify an appropriate statistical method for a given research situation, as well as for understanding, evaluating and carrying out statistical analyses and interpreting their results. In the problem proposed the students had to follow all the steps in a hypothesis test: 79.3% of them set correct or partially correct hypotheses; 64.1% of them correctly answered the two first questions and then computed the Chi-squared and *p*-value (with the help of software); 51.9% of students, in addition of completing the above steps, made a right decision and 43.5% of them correctly interpreted the results in the problem context and then answered correctly the four questions.

In the incorrect and partly correct responses we were able to identify several semiotic conflicts (Godino, 2002) where the interpretation of mathematical expressions by the students was different from what was expected by the lecturer. These semiotic conflicts involved either concepts/properties or procedures, as follows:

Semiotic conflicts involving concepts or properties

- Students exchanged the alternative and null hypotheses, an error described by Vallecillos (1994) and Chow (1996) (see IR.1 and IR.3 in part a). These students did not understand the differences between these two hypotheses: they could not understand that the null hypothesis is assumed to be true and determines the sampling distribution of the statistics, and that we cannot directly work with the statistical alternative hypothesis.
- Some students confused the statistics and the parameter and tried to establish the hypotheses in terms of the statistics (IR.2 in part a). Thus they failed to understand that while the probability distribution that models the values of a variable depends on the parameter values, while the statistics is computed from the sample data. Moreover in this response students did not notice that the *p*-value is a probability and therefore its value cannot be higher than 1.
- We also found some confusion between the *p*-value and the significance level, two concepts that, according to previous research, are particularly misunderstood. The confusion between these two concepts that was reported by Vallecillos (1994) was found in the response IR.4 in part b of the problem.
- Other students confused the acceptance and rejection regions as was visible in the incorrect responses in part c as well as in the responses IR.1 and IR.2 in part d of the

problem. These errors, also described by Haller & Kraus (2002), involve a misunderstanding of the logical reasoning behind statistical tests. According to Batanero (2000), this misunderstanding is due to the fact that, apparently, the formal structure of statistical tests is similar to that of proof by contradiction; however, there are fundamental differences between these two types of reasoning that are not always well understood by the students.

- The two hypotheses did not always covered the parametric space (see responses PR.3, IR.3 and IR.6 in part a); here again the student fails in understanding the logical reasoning in statistical tests, where the null hypothesis is the logical complement of the alternative statistical hypothesis.
- We also noticed confusion between a variable and its values (IR.4 in part a) or confusion between the number of variables and the number of values (IR.5 in part a).

Semiotic conflicts involving procedures

- In stating their hypotheses, part of the students confused the homogeneity test and independence test (responses IR.1 and IR.4 in part a) and, although in the teaching they received the lecturer distinguished between these two tests (as most statistics textbooks do), these students related 'independence' and 'Chi-square' when they were given a two-way table. As suggested by Seier (2006), after being accustomed to the idea of independence, it takes a while for students to later accept a more general framework and be aware that independence might not always be what we are curious about. It will be important to present the students with more examples for these situations, which are frequent in research; such as for example, testing the agreement of raters.
- Incorrect computation of the degrees of freedom in the Chi-squared distribution (responses IR.1, IR.3 in part b); this error was also observed in previous research by Alvarado (2007) and Olivo (2008); it also suggests the need to find ways to clarify to the students this concept, which is of great importance to modern statistical theory and is used in several important sampling distributions.

In the problem proposed, these students were asked to complete a modelling cycle. According to Henry (2001) and Chaput *et al.* (2011), modelling consists of describing an extra-mathematical problem in mathematical terms. This description leads to setting some hypotheses which are intended to simplify the situation. Next, the second step of the modelling process is translating the problem and the working hypotheses into a mathematical model in such a way that working with the model produces some possible solutions to the initial problem. The students translated the question (if the three words had identical emotional components) and the working hypotheses to statistical terms (they established their null and alternative hypotheses for the homogeneity). Consequently participants in our sample built and worked with statistical models (Chi-squared distribution, *p*-value and level of significance, critical and acceptance regions).

The third and final step consists of interpreting the mathematical results and relating these results to reality, in such a way that they produce some answers to the original problem. Although the majority of participants in our research correctly completed steps 1 and 2 in the modelling cycle, only 43.5% of them were capable of translating the statistical results they got to a response about the homogeneity of the words emotional component. That is, only this proportion of students could understand what the statistical results indicated about the original

problem and therefore, the remaining students failed to complete the last part of the modelling process.

Dantal (1997) suggests that in our classroom, we concentrate on step 2 ("the real mathematics" in the modelling cycle), since this is the easiest part to teach to our students. However all the steps are equally relevant for modelling and in learning mathematics, if we want our students to understand and appreciate the usefulness of mathematics. It is therefore very important that teachers of statistics try to develop the students' modelling ability.

Our results agree with Zieffler (2008), who suggests that statistics education research provides evidence about the existence of errors and faulty reasoning in students which can be quite difficult to correct in spite of good instructional methods and activities. A suggestion of the author is to use informal or formal methods of assessment to help students differentiate between their correct and incorrect ideas and procedures. In this sense, the problem posed in this research may be used in the classroom to discuss with the students the different steps needed in a homogeneity test and to warn students about possible errors in performing this procedure.

Acknowledgement: Project EDU2010-14947, grant FPU-AP2009-2807 (MCINN- FEDER) and BES-2011-044684 (MCINN-FEDER), and group FQM126 (Junta de Andalucía).

References

[1] Alacaci, G. (2004), Inferential statistics: Understanding expert knowledge and its implications for statistics education, *Journal of Statistics Education,* **12**(2), Online: www.amstat.org/publications/jse/v12n2/alacaci.html.

[2] Alvarado, H. (2007), *Significados del teorema central del límite en la enseñanza de la estadística en ingeniería* [Meaning of the central limit theorem in the teaching in engineering], Doctoral Thesis, Universidad de Granada, España.

[3] Batanero, C. (2000), Controversies around the role of statistical tests in experimental research, *Mathematical Thinking and Learning*, **2**(1-2), 75-98.

[4] Batanero, C. and C. Díaz (2006), Methodological and didactical controversies around statistical inference, *Actes des 38-ièmes Journées de Statistique*, CD-ROM, Paris: Societé Française de Statistique.

[5] Birnbaum, I. (1982), Interpreting statistical significance, *Teaching Statistics*, 4, 24–27.

[6] Castro Sotos, A. E., S. Vanhoof, W. Van den Nororgate, and P. Onghena (2007), Student´s misconceptions of statistical inference: A review of the empirical evidence form research on statistical education, *Educational Research Review*, **2**(2), 98-113.

[7] Chaput, B., J.-Cl. Girard, and M. Henry (2011), Modeling and simulations in statistics education. In Batanero C., G. Burrill, and C. Reading (Eds.) (2011), *Teaching Statistics in School Mathematics – Challenges for Teaching and Teacher Education. A Joint ICMI/IASE Study* (p. 85-95-83), Springer, New York.

[8] Chow, L. S. (1996), *Statistical significance: Rational, validity and utility*, Sage, London.

[9] Cohen, J. (1994), The earth is round (p<05), *American Psychologist*, **49**(12), 997-1003.

[10] Contreras, J. M., A. Estrada, C. Díaz, and C. Batanero (2010), Dificultades de futuros profesores en la lectura y cálculo de probabilidades en tablas de doble entrada [Prospective teachers' difficulties in computing probabilities from two-way tables]. In Moreno M., A. Estrada, J. Carrillo y T. Sierra (Eds.), *Investigación en Educación Matemática XIV* (p. 271-280), SEIEM, Lleida.

[11] Dantal, B. (1997), Les enjeux de la modélisation en probabilité [The betting of modelling in probability]. In Henry M. (Coord.), *Enseigner les probabilités au lycée* (p. 57-59), Commission Inter-IREM, Reims.

[12] Eco, U. (1979), *Tratado de semiótica general* [General semiotics], Lumen, Barcelona.

[13] Ellerton, N. (1996), Statistical significance testing and this journal, *Mathematics Education Research Journal*, **8**(2), 97-100.

[14] Falk, R. (1986), Misconceptions of statistical significance, *Journal of Structural Learning*, 9, 83-96.

[15] Falk, R. and C. W. Greenbaum (1995), Significance tests die hard: The amazing persistence of a probabilistic misconception, *Theory and Psychology*, **5**(1), 75-98.

[16] Godino, J. D. (2002), Un enfoque ontológico y semiótico de la cognición matemática [An ontological and semiotic approach to mathematical cognition], *Recherches en Didactique des Mathématiques,* **22**(2 et 3), 237-284.

[17] Godino, J. D., C. Batanero, and V. Font (2007), The onto-semiotic approach to research in mathematics education, *ZDM, The International Journal on Mathematics Education*, **39**(1-2), 127-135.

[18] Hackett, R. (2010), Contrasting cases: the "b versus c" assessment tool for activating transfer. In Reading C. (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics*, Ljubljana, Slovenia, International Statistical Institute, www.stat.auckland.ac.nz/~iase/publications.php.

[19] Haller, H. and S. Krauss (2002), Misinterpretations of significance: A problem students share with their teachers?, *Methods of Psychological Research*, **7**(1), Online: www.mpr-online.de/.

[20] Harlow, L. L., S. A. Mulaik, and J. H. Steiger (1997), *What if there were no significance tests?*, Mahwah, NJ: Lawrence Erlbaum Associates.

[21] Harradine, A., C. Batanero, and A. Rossman (2011), Students and teachers' knowledge of sampling and inference. In Batanero C., G. Burrill, and C. Reading (Eds.), *Teaching statistics in school mathematics. Challenges for teaching and teacher education*, Springer.

[22] Henry, M. (2001), Notion de modèle et modélisation dans l'enseignement. [Notion of model and modelling in teaching]. In Commission Inter-IREM Statistique et Probabilités, *Autour de la modélisation en probabilités* (p. 149-159), PUFC, Besançon, France.

[23] Lecoutre, B. and M. P. Lecoutre (2001), Uses, abuses and misuses of significance tests in the scientific commumity: Won't the Bayesian choice be unavoidable?, *International Statistical Review*, **69**(3), 399-417.

[24] Levin, J. R. and D. H. Robinson (1999), Further reflections on hypothesis testing and editorial policy for primary research journals, *Educational Psychological Review*, 11, 143-155.

[25] Liu, Y. and P. W. Thompson (2009), Mathematics teachers' understandings of proto-hypothesis testing, *Pedagogies*, **4**(2), 126-138.

[26] Olivo, E. (2008), *Significados de los intervalos de confianza para los estudiantes de ingeniería en México* [Meaning of the confidence intervals for Mexican students of engineering], Doctoral Thesis, Universidad de Granada, España.

[27] Robinson, D. H. and J. T. Levin (1997), Reflections on statistical and substantive significance, with a slice of replication, *Educational Researcher*, **26**(5), 21-26.

[28] Seier, E. (2006), Influence of consulting in the selection of topics when teaching statistics. In Rossman A. and B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Bahia, Brazil: International Statistical Institute and International Association for Statistical Education, Online: www.stat.auckland.ac.nz/~iase/publications.

[29] Vallecillos, A. (1994), *Estudio teórico-experimental de errores y concepciones sobre el contraste estadístico de hipótesis en estudiantes universitarios* [An empirical-theoretical study of errors and conceptions on statistical tests in university students], Doctoral Thesis, Universidad de Granada, España.

[30] Vallecillos, A. and C. Batanero (1997), Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios [Concepts activated in statistical tests and its understanding by university students], *Recherches en Didactique des Mathematiques*, 17, 29-48.

[31] Vera, O., C. Díaz, and C. Batanero (2011), Dificultades en la formulación de hipótesis estadísticas por estudiantes de Psicología [Difficulties in formulating statistical hypotheses by Psychology students], *Unión*, 27, 41-61.

[32] Walker, H. M. (1940), Degrees of freedom, *Journal of Educational Psychology*, **31**(4), 253-269.

[33] White, A. L. (1980), Avoiding errors in educational research. In Shumway R. J. (Ed.), *Research in mathematics education* (p. 47-65), Va: National Council of Teachers of Mathematics, Reston.

[34] Wilkinson, L. (1999), Statistical methods in psychology journals: Guidelines and explanations, *American Psychologist*, 54, 594-604.

[35] Zieffler, A. (2008), What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature, J*ournal of Statistics Education*, **16**(2),
Online: www.amstat.org/publications/jse/v16n2/zieffler.html.