Journal of Statistics Education

How confident are students in their misconceptions about hypothesis tests?

Ana Elisa Castro Sotos
Stijn Vanhoof
Wim Van den Noortgate
Patrick Onghena
Centre for Methodology of Educational Research,
Katholieke Universiteit Leuven

Journal of Statistics Education Volume 17, Number 2 (2009), www.amstat.org/publications/jse/v17n2/castrosotos.html

---

Key Words: "Confidence"; "University Students"; "Misconceptions *p*-value"; "Misconceptions Significance Level".

Abstract

Both researchers and teachers of statistics have made considerable efforts during the last decades to re-conceptualize statistics courses in accordance with the general reform movement in mathematics education. However, students still hold misconceptions about statistical inference even after following a reformed course. The study presented in this paper addresses the need to further investigate misconceptions about hypothesis tests by (1) documenting which misconceptions are the most common among university students of introductory courses of statistics, and (2) concentrating on an aspect of research about misconceptions that has not yet received much attention thus far, namely the confidence that students have in their misconceptions. Data from 144 college students were collected by means of a questionnaire addressing the most common misconceptions found in the literature about the definitions of hypothesis test, *p*-value, and significance level. In this questionnaire, students were asked to select a level of confidence in their responses (from 0 to 10) for each item. A considerable number of participants seemed to hold misconceptions and lower levels of concept-specific self-perceived efficacy were found to be related to misconceptions more than to the correct answers. On average, students selected significantly lower levels of confidence for the question addressing the definition of the significance level than for the other two items. Suggestions for further research and practice that emerge from this study are proposed.

1. Introduction

During the last decades, students' most common misconceptions about statistical inference have been extensively documented (for a review see Castro Sotos, Vanhoof, Van den Noortgate, and Onghena 2007) and the severe impact of these errors on the research community has been stressed (Batanero 2000; Gliner, Leech, and Morgan 2002; Haller and Krauss 2002). In response to the persistence of the misconceptions, educational researchers and practitioners have initiated and promoted a thorough reform for teaching statistics within the broader framework of the reform movement in mathematics education that has taken place during the last decades (NCTM 1989, 2000).

The reform in statistics education has been based on updating the three pillars for statistics courses: *Content*, *pedagogy*, and *technology* (Moore 1997). First, the reform has promoted a shift of these courses' focus from a mathematical perspective, centered on computation and procedures, towards a new emphasis on the ideas of statistical *reasoning*, statistical *thinking*, and statistical *literacy* (Ben-Zvi and Garfield 2004). This shift has meant a transformation of probability-based statistics courses into data-based courses. Second, parallel to laying this new foundation, the reform movement has firmly encouraged the use of, among other things, real data and group assignments as efficient didactical measures for teaching statistics. These activities have been expected to improve students' collaborative and communicative skills through statistical discussions (e.g., Garfield 1993; Keeler and Steinhorst 1995; Giraud 1997; Derry, Levin, Osana, Jones, and Peterson 2000). Finally, the importance of integrating technology in the statistics classroom in synergy with the new data-oriented perspective and revised pedagogies has been equally emphasized (e.g., Schuyten, Dekeyser, and Goeminne 1999; Dutton and Dutton 2005). Specific software such as Sampling Sim (delMas 2001) or Fathom (Finzer 2005) has been developed to help students understand the ideas behind statistical processes by means of, for example, simulation.

Despite the considerable efforts made by educational researchers and statistics instructors, students have been found to still hold misconceptions after following a reformed course (González and Birch 2000; Chance, delMas, and Garfield 2004; Scheines, Leinhardt, Smith, and Cho 2005; delMas, Garfield, Ooms, and Chance 2007). In order to enlarge the research base for the construction of powerful learning environments for statistics, new research studies should focus on two aspects. First, they should continue developing and improving methods for teaching statistics and assessing students' statistical reasoning (Garfield 2001; delMas et al. 2007). Second, they should further investigate university students' conceptual understanding in statistics (Shaughnessy 2007) and, more specifically, misconceptions (Innabi 1999). Additional research about misconceptions is necessary, as argued by Rossman and Chance (2004), to learn how to *recognize*, *diagnose*, and *address* these errors in order to more efficiently help students learn statistics. In fact, gathering as much information as possible about students' misconceptions can be very helpful for the development of new assessment techniques that acknowledge those common errors. In particular, special attention should be paid to misconceptions concerning hypothesis tests, given the widespread use of statistical tests in all areas of research (Haller and Krauss 2002).

The present study addresses this second need of further investigating misconceptions about hypothesis tests in two ways. First, we document which misconceptions pertaining to the information that a statistical tests provides are the most common among university students of introductory courses of statistics. And second, we pay attention to an aspect of misconceptions' research that has not yet received much attention; namely the confidence that students have in their misconceptions.

Before describing in more detail these two research issues, and for the sake of clarity, we specify here our position with regard to the discussion that has been raised about a proper definition for the term *misconception*, and how it should, or should not, be distinguished from terms such as *preconception*, *misunderstanding, or misinterpretation* (Guzzetti, Snyder, Glass, and Gamas 1993; Smith III, diSessa, and Roschelle 1993). In this manuscript, we follow Cohen, Smith, Chechile, Burns, and Tsai (1996), using a broad characterization of the concept of *misconception* that embraces all those ideas, this term referring to any sort of fallacy, misunderstanding, misuse, or misinterpretation of a concept, provided that it results in a documented systematic pattern of error.

1.1 Misconceptions About Hypothesis Tests

Getting students to make sense of hypothesis tests has proven to be a very difficult task for statistics instructors because of the persistency and deepness of the misconceptions held by learners (Brewer 1985; Daniel 1998; Kirk 2001), even after years of training (Falk 1986). In fact, Mittag and Thompson (2000), Gordon (2001), and Lecoutre, Poitevineau, and Lecoutre (2003) convincingly showed that even statisticians are not immune to misconceptions about hypothesis tests. For more than twenty years, misconceptions and misuses regarding hypothesis tests have been documented (an historical summary of bad practices can be found in Daniel 1998).

The three most difficult aspects of testing statistical hypothesis are (Batanero, 2000): Understanding the concept of a hypothesis test, interpreting a *p*-value, and interpreting the significance level a. Empirical confirmation of several misconceptions related to these ideas can be found for example in Williams (1998), Vallecillos (2000), and Haller and

Krauss (2002), or, most recently, in delMas et al. (2007), who provided new evidence of misconceptions from a very large sample of students across the United States.

Based on the ideas and evidence from this type of publication, we decided to investigate the errors listed below, which seemed to be the most common and persistent ones in the literature (see also Falk and Greenbaum 1995; Gliner et al. 2002). These misconceptions pertain to the definitions of hypothesis test (errors called *H1* and *H2* below), *p*-value (called *p1* to *p6*), and significance level a (a*1* to a*5*):

Considering the hypothesis test as a mathematical (logical) proof of one of the two hypotheses (H1). Holding this conception means assuming that, just as any mathematical procedure, the results of a hypothesis test are deterministic, therefore believing that the null (or the alternative) hypothesis has been fully proven to be true or false with a 100% certainty.

Considering the hypothesis test as a probabilistic proof by contradiction (Illusion of attaining improbability) of one of the two hypotheses (H2). This error is a consequence of the similarity in the formal structures of the reasoning in hypothesis testing and the mathematical proof by contradiction, which is based on the logical *modus tollens* method. As can be seen in Table 1, the misconception arises when the modus tollens method is applied to the results of a hypothesis test. This equivalence to the mathematical proof by contradiction does not work for hypothesis tests: While a contradiction disproves the premise from which it is drawn, a low probability event does not make the premise from which it is drawn improbable. The difference with the previous misconception (*H1*) is that here the null (or the alternative) hypothesis is believed to have been proven *probably* true or *probably* false, as opposed to the 100% certainty mentioned above.

Table 1. Modus Tollens and the Illusion of Probabilistic Proof by Contradiction

Statements	Modus tollens	Example	Illusion
Premise 1	$p \rightarrow q$	If it is raining there are clouds	If H_0 is true there is a high probability that the *p*-value is large
Premise 2	q^c	There are no clouds	The *p*-value is small
Conclusion	p^c	It is not raining	H_0 is improbable

Misconceptions arising from the spread confusion of switching the two terms in the conditional probabilities that define the ideas of p-value and significance level a.

p1 The *p*-value is the probability of the null hypothesis assuming the same (or more extreme) data.
a1 The significance level a is the probability of the null hypothesis assuming its rejection.

Misconceptions occurring when simply ignoring the conditioning event.

p2 The *p*-value is the probability of obtaining the same (or more extreme) data.
a2 The significance level a (or 1-a) is the probability of rejecting the null hypothesis.

Misconceptions resulting from the combination of the two previous confusions.

p3 The *p*-value is the probability of the null (or alternative) hypothesis.
a3 The significance level a is the probability of the null (or alternative) hypothesis.

Additional misconceptions documented in the literature.

p4 The *p*-value is the probability of making an error when rejecting the null hypothesis.
p5 The *p*-value indicates how big is the distance between groups under test.
p6 The *p*-value indicates how small is the distance between groups under test.
a4 Statistically significant results for a prove that the null hypothesis is improbable

a5 Statistically significant results for a prove the falseness of the null hypothesis

1.2 Students' self-perceived efficacy

For the present study, we considered *confidence* as students' self-(perceived) efficacy (Bandura 1977) in answering the items about hypothesis tests. There is a profuse amount of literature and abundant empirical support for the concept of self-efficacy (for a review of the different theories see Eccles and Wigfield 2002). In fact, self-efficacy has been emphasized as a central factor affecting academic proficiency (Stajkovic and Luthans 1998), also for statistics courses (Finney and Schraw 2003).

Specifically, "confidence in statistics" is the focus of one of the subscales (*Cognitive Competence* scale) from the well-known Survey of Attitudes Toward Statistics (Schau, Dauphinee, Del Vecchio, and Stevens 1999). However, this scale seems to address the idea of confidence as a broader academic self-efficacy notion referring to the domain in general. This type of "general confidence in a field" is not as strongly related to academic performance as self-efficacy beliefs that are more task and situation-specific (Pajares 1996). For example, Finney and Schraw (2003) found no relation between self-efficacy and performance in their study and discussed whether it could have been due to their use of a measure that was not task-specific (but domain-specific) and would decontextualize self-efficacy judgments. According to these authors, "[…] the closer the correspondence between the task and self-efficacy assessment, the better the prediction of performance on the task […]" (Finney and Schraw 2003, p. 163). Moreover, as supported by social cognitive theory, the relation between self-efficacy and performance is the strongest when it is measured at the optimal level specific to the task (Choi 2005). Correlations between task-specific self-efficacy beliefs and academic performance in those tasks have been found to vary between r=0.49 and r=0.70 (Pajares 1996; 0.38 for the unbiased effect size estimate of the meta-analysis by Multon and Brouwn 1991). These correlations confirm the stronger relation with performance of specific-type of self-efficacy measures as opposed to more general self-efficacy beliefs.

We could find sparse previous research in the statistics education literature about the relation between confidence in a specific concept and the correctness of such concept; and those studies that looked into it have reported different results. For example, Allen, Reed, Rhoads, and Terry (2006) used confidence ratings about introductory statistics concepts and found a positive trend between correct responses and confidence across items. On the other hand, Chance et al. (2004) could not find any relevant relation between students' confidence in their responses to items about sampling distributions and the correctness of those answers.

With the intention of shedding new light on this relationship for statistics concepts, we focus our interest for this study on the connection between the correctness of participants' answers (misconceptions present vs. correct answer) and their confidence in those answers for each of the items about hypothesis tests.

2. Methodology

To address our twofold goal, we administered a questionnaire including three items about hypothesis tests to 144 university undergraduates following introductory statistics courses. In this questionnaire, students could also indicate the confidence in their specific answers.

The questionnaire was constructed inspired by items used in previous research (e.g., Falk 1986; Vallecillos 2000), the items' database from the ARTIST website (https://app.gen.umn.edu/artist/), and our own experience. The instrument was made up of five multiple-choice items of which three addressed misconceptions concerning hypothesis tests (see Appendix for the items). Students were asked to select the correct answer for each item; however, they often picked more than one response.

The first of the three target items asked students to select the correct definition of a hypothesis test out of six possible statements. The first two phrases represented misconception *H1* for the null and alternative hypotheses respectively. The following two statements corresponded to misconception *H2* (again for the null and alternative hypotheses respectively). The fifth option was the correct one, and finally, the sixth phrase referred to a correct probability definition but applied to the alternative hypothesis.

The second item addressed the concept of *p*-value and presented participants with seven statements of which they had to select the one that correctly defined this notion. The order of statements in relation to the misconceptions described above was the following: *p3*, *p1*, *correct*, *p2*, *p4*, *p5*, and *p6*.

Finally, the third question dealt with the possible misinterpretations of the significance level a in the same way as the previous two items, students had to select the correct option out of a list with only one correct answer and the misconceptions mentioned above in the following order: *a4*, *a5*, *a2*, *correct*, *a1*, and *a3*.

With the purpose of exploring the relation between the manifestation of the different misconceptions and students' confidence, just below each item we required participants to select how confident they were about their specific answer on a 10-point Likert-scale (*0*=0% confident my answer is correct, *10*=100% confident my answer is correct, see Appendix for the complete items).

The questionnaire was distributed in January 2007 to 144 Spanish university students (95 females, 48 males, 1 not responded), from three different disciplines (Mathematics 21%, Medicine 52%, and Business studies 23%, 4% not responded), who were taking introductory statistics courses that covered the topic of hypothesis tests at the Complutense University (Madrid, Spain). During their statistics course, students were trained to interpret the process and resulting *p*-value of a hypothesis test. Half of the participants were in their first year. Before the present course, most participants (88%) had taken no university statistics course or only a course where they learned about descriptive statistics, forming a fairly homogeneous group as far as statistical background was concerned. The questionnaire was voluntarily completed and handed in during one of the last lessons of the semester in the presence of the teacher, as an anonymous exercise.

Regarding the analysis of the data, we used descriptive statistics to provide an overview of the results on the appearance of misconceptions and intensity of students' confidence in their answers. In addition, we used two-sample tests for proportions to compare responses from students with and without statistical background, and non-parametric tests for the comparison of participants' confidence in their responses between the three items. Furthermore, in order to graphically observe the distribution of students' confidence for the different answers and items, we used box-plots, which also helped us to get a first impression of the possible relation between level of confidence and correctness of the answer. To further investigate that relation, we performed two-sample *t*-tests to look for significant differences between levels of confidence according to correctness of answer (completely correct answer vs. presence of misconceptions). In addition, we calculated point-biserial correlations between levels of confidence and presence/absence of misconceptions for the three items. The results of these analyses are presented in the following section.

3. Results

3.1 Presence of Misconceptions

A summary of the number of students who selected each of the available statements concerning the definition of hypothesis tests (first target item) can be found in Table 2. More than half (62%) of the 144 participants marked the correct option for this item. However, this percentage decreases when we restrict the criteria to selecting the correct statement *plus* not selecting any of the other alternatives (completely correct answer, no misconception present, 56%). As can be seen, the two major misconceptions for the concept of hypothesis test were just as popular among our students: 20% of them believed that a hypothesis test is a mathematical (logical) proof of the null hypothesis, and 19% that it is a probabilistic proof by contradiction.

For all three options (*H1*, *H2*, and *correct*), most participants seemed to recognize that the evidence when performing a hypothesis test refers to the null hypothesis that will be (or not) refuted. Only small percentages of students (1%, 3%, 6% respectively) confused the null with the alternative hypothesis for the different statements.

Table 2. Percentages of selected responses for the definition of hypothesis test (total *n*=144)

Option	"A hypothesis test is…	Nr. Students	Percentage
H 1	…a proof of H_0 (or) or its falseness"	29	20%
	…a proof of H_a (or) or its falseness"	2	1%
H 2	…a proof of the probability or improbability of H_0"	27	19%
	…a proof of the probability or improbability of H_a"	5	3%
Correct	…an evaluation of the evidence in favor or against H_0 in the data"	89	62%
Confusing H_0 and H_a	…an evaluation of the evidence in favor or against H_a in the data"	8	6%
Completely correct answer: Selecting the correct option only		80	56%

Note: For a correct interpretation of the percentages in the table, take into account that, although they were told to select *the* most correct option, students often picked more than one response and therefore these percentages do not sum up to 100% neither does the number of students sum up to 144.

Table 3 presents the summary of selected responses for the definition of a *p*-value in our study (second target item). The correct characterization was selected by 52% of the participants and it is here also the case that the percentage of students who not only found the correct option but also left all other incorrect statements out of their answer (completely correct answer) is smaller, namely 46%.

As can be observed in the table, participants' most popular misconception was that of *p4* (16%), stating that the *p*-value is the probability of making an error when rejecting the null hypothesis. This error can be seen as a derivation of *p1* if "obtaining the same or more extreme data" is considered as automatically rejecting the null hypothesis, and therefore the null hypothesis being true means making an error. In both cases, *p1* and *p4*, the misconception results from inverting the ideas in the conditional probability, considering that the action (rejecting the null hypothesis or, equivalently, obtaining extreme data) is the conditioning event for whether the null hypothesis is true or not. If, aside from the information in the table about the selection of individual statements, we calculate the amount of participants who chose *p1*, *p4*, or both, we obtain a percentage of 23% (33 out of 144). Hence, we could summarize by saying that the most common problem that participants showed was that of erroneously inverting the terms in the conditional probability that defines the concept of *p*-value.

Table 3. Percentages of selected responses for the definition of the *p*-value (*n*=144)

Option	"Given a *p*-value of 0.01…	Nr. Students	Percentage
p3	… the probability of the null hypothesis is 0.01"	17	12%
p1	… the probability of the null hypothesis assuming the same or more extreme data is 0.01"	13	9%
Correct	… the probability of obtaining the same or more extreme data assuming the null hypothesis is true"	75	52%
p2	… the probability of obtaining the same or more extreme data is 0.01"	10	7%
p4	… the probability of making an error when rejecting the null hypothesis is 0.01"	23	16%
p5	…the difference between the groups is big"	11	8%
p6	…the difference between the groups is small"	15	10%
Completely correct answer: Selecting the correct option only		66	46%

Note: For a correct interpretation of the percentages in the table, take into account that, although they were told to select *the* correct option, students often picked more than one response and therefore these

percentages do not sum up to 100% neither does the number of students sum up to 144.

The summary for the different answers to the item on the definition of significance level (third target item) can be found in Table 4. As in the previous two items, the percentage of students opting for a completely correct answer is smaller than that of simply selecting the appropriate definition of the significance level besides some other alternative(s). In this case, the difference is, however, smaller: 40% versus 42%.

For the significance level, responses were more spread among the different statements than in the previous two items. With 17% of selection, a*2* and a*3* were the most common misconceptions for our students, closely followed by a*4* and a*1* (10% selection). It is interesting to notice that both a*2* and a*3* correspond to simple probabilities (not conditionals) as opposed to the more elaborated ideas presented by a*4* and a*1*. However, a*5*, which was one of the most straightforward options, happened to be very unpopular.

Table 4. Percentages of selected responses for the definition of the significance level

Option	"If the results of a test are statistically significant for a level $a = 0.05$ …⋯	Nr. Students	Percentage
a4	… the null hypothesis is proven to be improbable"	14	10%
a5	… the null hypothesis is proven to be false"	9	6%
a2	…the probability of rejecting the null hypothesis is of 95%"	24	17%
Correct	… the probability of rejecting the null hypothesis assuming the null hypothesis is true"	60	42%
a1	…the probability of the null hypothesis assuming that it is rejected is of 5%"	15	10%
a3	…the probability of the null hypothesis is of 5%"	24	17%
Completely correct answer: Selecting the correct option only		57	40%

Note: For a correct interpretation of the percentages in the table, take into account that, although they were told to select *the* correct option, students often picked more than one response and therefore these percentages do not sum up to 100% neither does the number of students sum up to 144.

Our participants came from different groups, differing in the majors that were followed, in the textbooks and in the instructors. However, no significant differences between the groups were found with regard to the presence of the misconceptions and the patterns shown above (Tables 2, 3, and 4).

When looking at possible connections between the items we observe three interesting phenomena. First, when comparing the results of the first and the third questions, we find that students chose misconceptions *H1* and *H2* more often than a*4* and a*5* respectively (20% vs. 10% and 19% vs. 6%), whereas they could be seen as equivalent. *H1* considers a test as a mathematical proof and a*5* states that significant results for a=0.05 mean that the null hypothesis has been proven to be false. Similarly, *H2* considers the test as a probabilistic proof while a*4* affirms that significant results for a=0.05 mean that the null hypothesis has been proven to be improbable. In fact, it can be seen as a contradiction that only three students selected both options *H1* for the first item and a*5* for the third, whereas all the other students choosing *H1* for the first item opted for a different option in the third question. Likewise, only 4 participants chose *H2* and a*4*.

Secondly, although a*1* and *p1* (and/or *p4*) are highly similar, percentages differ considerably here as well. Just 10% of the participants selected a*1* whereas 23% agree with *p1*, *p4*, or both. a*1* states that significant results for a=0.05 indicate that the probability of the null hypothesis assuming that it is rejected is of 5%, while *p1* and *p4* result also from the inversion in the terms of the conditional probability that defines the *p*-value (see above). Again just a few

students (5) selected *p1* (or *p4* or both) and a*1*.

The third interesting fact arises when comparing the misconceptions *p3* and a*3*. In this case, the results indicate comparable percentages of selection (12% and 17%) for these two statements than can be seen as similar. Both *p3* and a*3* address the concept in question (*p*-value and significance level respectively) as the single probability of the null hypothesis. However, once more, only a very small number of participants (3) selected *p3* for the second item and a*3* for the third.

For all three items, we compared the correctness of the answers between students with and without statistical background (students who took at least another statistics course before vs. students who took none). We found that students without previous training score significantly better than those with experience do for the items about the definition of hypothesis tests and about the *p*-value (see Table 5). For the first question, 69% of the students without statistical background provided a completely correct answer, versus 44% of the students who had taken at least another statistics course before. For the second item, the difference between these percentages was even larger (73% versus 18%). Table 5 shows the results of the tests for proportions of completely correct answers by type of background that confirm these differences for the first two items to be significant.

Table 5. Test for proportions of completely correct answers by backgrounds

H_0	Z	Interval for difference (95%)	*p*
Freq. Comp. Correct Answer HT(No Background) = Freq. Comp. Correct Answer. HT (Background)	2.81	(0.077, 0.405)	**0.0050**
Freq. Comp. Correct Answer *p*-value (No Background) = Freq. Comp. Correct Answer. *p*-value (Background)	6.39	(0.414, 0.694)	**<.0001**
Freq. Comp. Correct Answer a (No Background) = Freq. Comp. Correct Answer a (Background)	-1.05	(-0.258, 0.077)	0.2924

Note: Participants who did not answer any of the three questions or did not provide information about their statistical background were not included in this analysis (*n*=133).

Lower percentages of correct answers were found for the third item as compared to the percentages for the first two questions (see Tables 4, 2, and 3: 42% vs. 62% and 52%, the only significant difference is between the first and the third item with *p*=0.0047). In addition, only for the third question (about the significance level) we found no significant difference between experienced and novice students. It is difficult to find the specific reasons for this since we do not have enough detailed information regarding the students' background. It could be the case that some students who selected "I have taken a statistics course before" did not yet take another course, but were repeating the introductory course since they did not pass the first time. In addition, the content of the previous course and the teacher can be relevant factors playing a role in these differences. However, because we would expect that these factors affect all three items to the same degree, but still the results for the third item seem to follow a different pattern than the results for the first two, we take this finding as an indication of an underlying difference between the first two items and the third one, that is further explored by means of the analysis of students' confidence in the next section.

3.2 Students' confidence

As described in Section 2, following each item, students had to select their confidence in the response provided on a 10-point Likert-scale (*0*=0% confident my answer is correct, *10*=100% confident my answer is correct). A summary of participants' confidence in their answers can be found in Table 6. Note that not all the students selected a confidence level for their answers for all items.

Table 6. Summary of confidence level for each item

Item	*n*	Mean Confidence	Standard Deviation	Min. Selected	Max. Selected
Hypothesis Test	142	6.21	2.58	0	10
p-value	140	5.66	2.96	0	10
Significance level a	134	4.69	2.79	0	10

In line with some of the results presented above (percentage of correct answers and difference between experienced and novice students), we again observe a different pattern of results for the third question (about the significance level) when looking at students' confidence. Not only was the number of students providing a confidence level lower for the third item than for the other two (n=134 vs. n=142 and n=140; no significant differences), but also those participants who did report a certain level for all three questions were less confident on average about their answers to the third one than to the other two (see Table 7).

Table 7. Summary of students providing a confidence level for all three items

Item	*n*	Mean Confidence	Standard Deviation	Min. Selected	Max. Selected
Hypothesis Test	132	6.21	2.56	0	10
p-value	132	5.63	2.98	0	10
Significance level a	132	4.70	2.81	0	10

To compare participants' confidence in their responses to the first, second, and third items, we constructed three contrast-variables for the difference between these (paired) confidences for each student (student's confidence on item 1 – same student's confidence on item 3, student's confidence on item 2 – same student's confidence on item 3, and student's confidence on item 1 – same student's confidence on item 2). Because we had to reject the hypothesis that these new variables were normally distributed (p-values for Shapiro-Wilk and Kolmogorov-Smirnov tests for the three variables were all <0.01), we performed non-parametric tests for the equality to zero of the median for each variable (see Table 8). These tests confirmed a significant difference in median confidence between the question about the hypothesis test and significance level (symmetric distribution: p-value for the Sign Test $p<0.0001$), as well as between the confidences in the item about the p-value and the item about the significance level (symmetric distribution: p-value for the Sign Test $p=0.001$). On the other hand, the difference in median confidence between the questions about the hypothesis test and the p-value was not significant (asymmetric distribution: p-value for the Wilcoxon Signed Rank Test $p=0.2229$).

Table 8. Medians for Contrast (Paired) Confidences

Variable	*n*	Mean	Standard Deviation	Min	Max
Confidence Item 1 – Confidence Item 3	132	1.51	2.64	-6	9
Confidence Item 2 – Confidence Item 3	132	0.92	3.13	-7	9
Confidence Item 1 – Confidence Item 2	132	0.58	2.39	-4	8

This phenomenon, together with the lower number of correct answers (see above) and the non-significant difference in the performance of experienced and novice students (see Table 5) for the third item as opposed to the other two, could be revealing again that the third question about the concept of significance level was slightly more difficult for our participants or, at least, more confusing.

For the purpose of investigating whether there exists any identifiable relation between the manifestation of the misconceptions and the confidence of participants in them, we compared the average confidence of students providing a completely correct answer with that of students committing some misconception. For all three questions, the average confidence of those students who provided a completely correct answer was higher than for those who selected at least one of the misconceptions (see Figure 1).

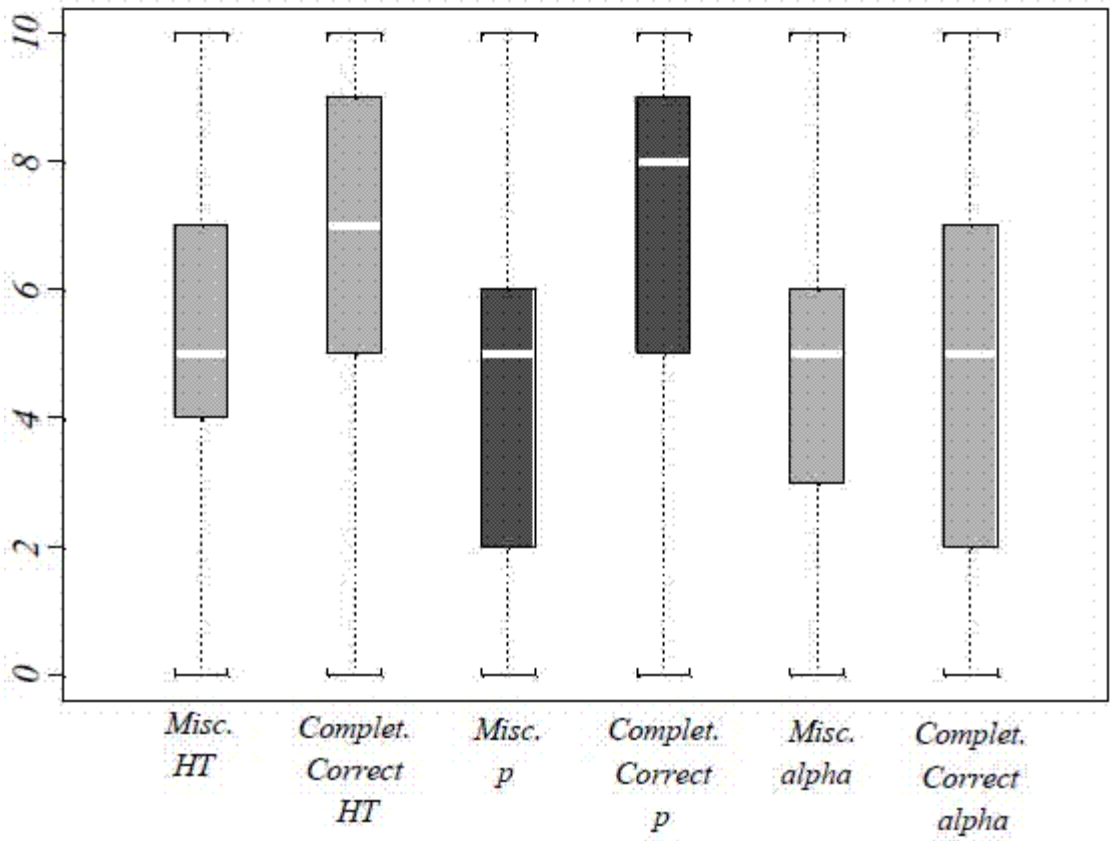

Figure 1. Confidence of students with correct answer vs. students with misconception for all three items

Note: Sample sizes of the groups are HT item: Completely correct *n*=75, Misc *n*=57; *p*-value item: Completely correct=62, Misc *n*=70; a item: Completely correct *n*=57, Misc *n*=75. The symbol "*a*" has been substituted by the word "alpha" in the horizontal-axis.

These differences were statistically significant for the items concerning both the hypothesis test definition and the *p*-value, but not for the question about the significance level a (see Table 9); which again points at a higher difficulty and/or at least more uncertainty in the answers to this question.

Table 9. Two-sample *t*-tests for Confidence classified by type of answer (completely correct vs. misconception)

Item	Perfect Answer	Mean Confidence	Stand. Dev.	DF	*t*-value	*p*-value
Hypothesis Test	No	5.33	2.36	130	-3.60	**0.0005**
	Yes	6.88	2.51			
p-value	No	4.47	2.85	130	-5.19	< **.0001**
	Yes	6.94	2.57			
Significance level a	No	4.65	2.65	130	-0.24	0.8112
	Yes	4.77	3.02			

When looking more closely at the confidence of those students who selected one and only one statement (those putting all their confidence in just one option), we observe again a difference between the first two items and the one about the significance level. Comparing the width of the box-plots (representing sample sizes) in Figures 2 to 4 we observe that in the case of the items about the definition of a hypothesis test and a *p*-value, a large part of the students gave the correct answer and were relatively confident about their answer. For the question about the significance level, the responses are much more spread and again, the idea arises that this question was more confusing for the students than the other two.

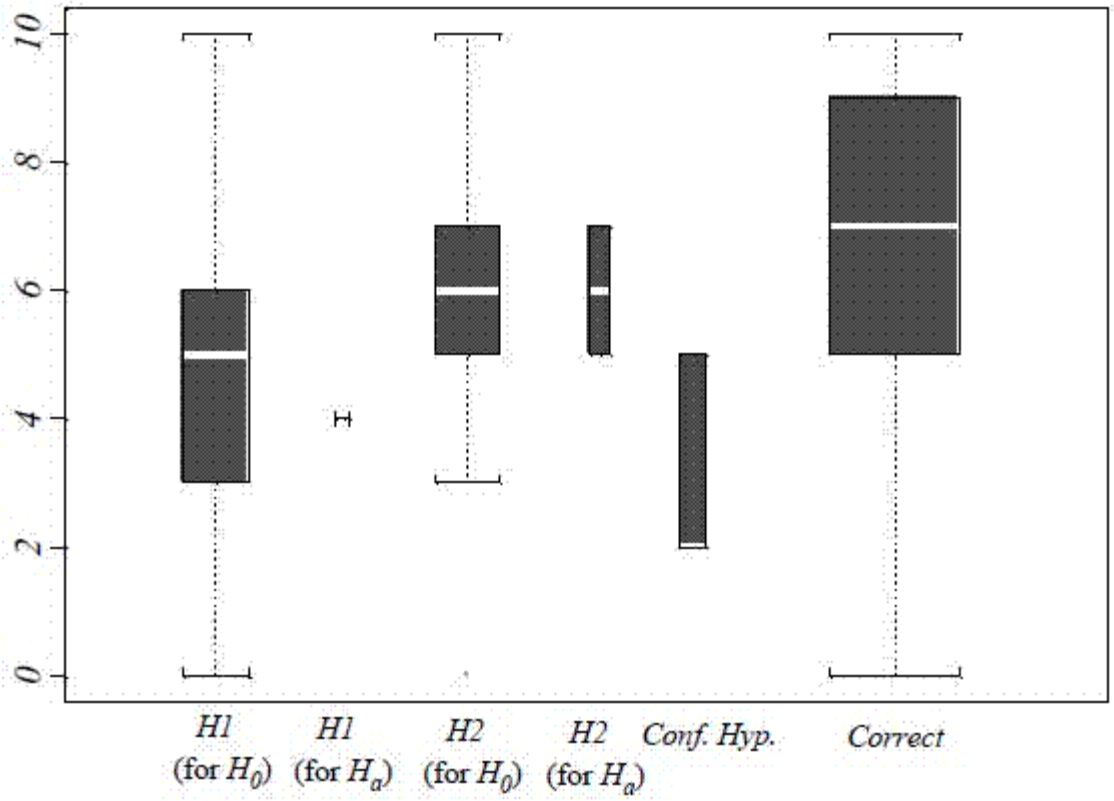

Figure 2. Confidence of students selecting each (and only) option for the first item.

Note: The width of the boxes is proportional to the number of students selecting each (and only) option. *H1* (for H_0): 19, *H1* (for H_a): 1, *H2* (for H_0): 18, *H2* (for H_a): 2, *Conf. Hyp*: 3, *Correct*: 73.

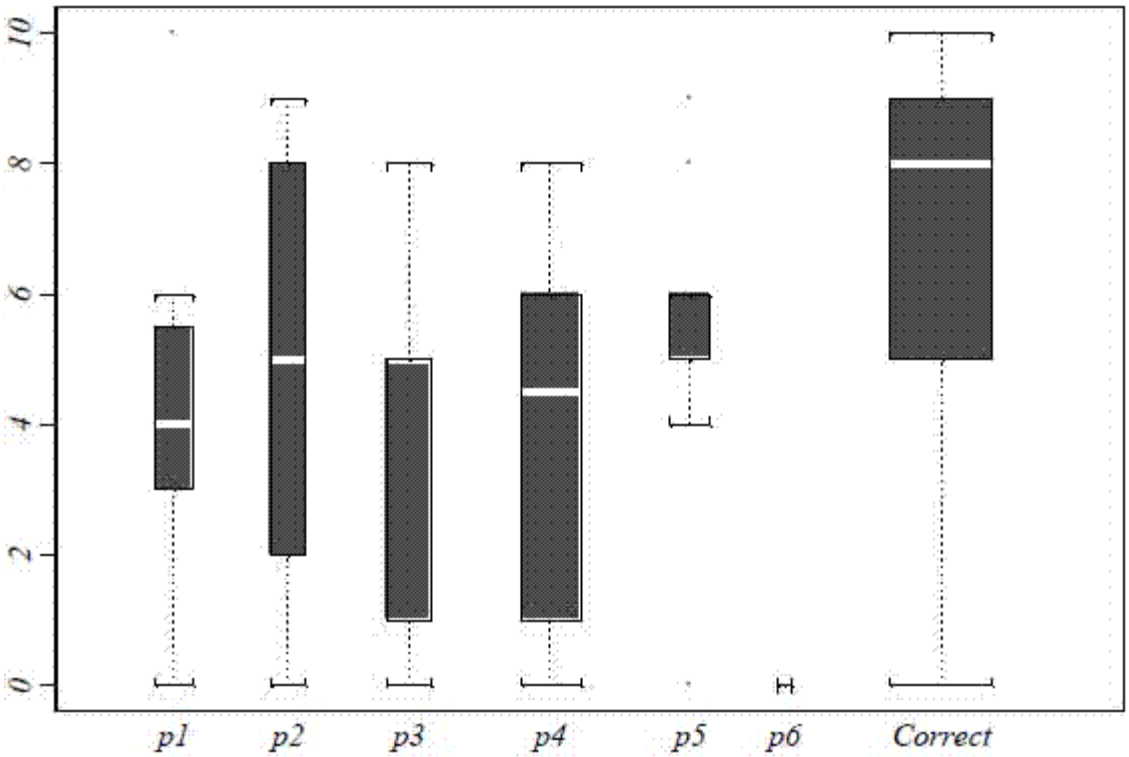

Figure 3. Confidence of students selecting each (and only) option for the second item.

Note: The width of the boxes is proportional to the number of students selecting each (and only) option. *p1*: 8, *p2*: 7, *p3*: 11, *p4*: 18, *p5*: 9, *p6*: 1, *Correct*: 59.

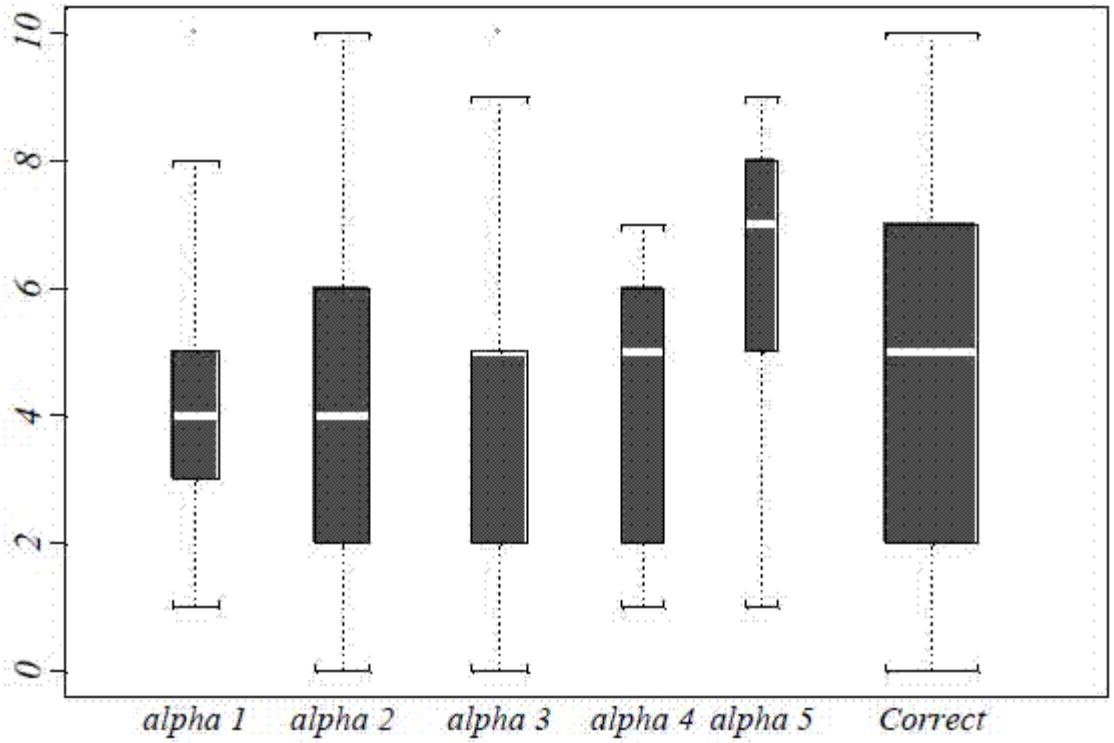

Figure 4. Confidence of students selecting each (and only) option for the third item.

Note: The width of the boxes is proportional to the number of students selecting each (and only) option. *a1*: 13, *a2*: 19, *a3*: 19, *a4*: 11, *a5*: 6, *Correct*: 53. The symbol *a* has been substituted by the word "alpha" in the x-axis.

Figures 2 and 3 and the *t*-tests in Table 9 seem to substantiate the trend found by Allen et al. (2006) that higher confidence levels coincide with correct answers more than with incorrect ones. The point-biserial correlations, equivalent to the Pearson's correlation for a continuous variable (Confidence) and a dichotomous variable (Presence of misconception), for the three target questions were r_{pb}=0.36, r_{pb}=0.53, and r_{pb}=0.02 respectively. These correlations confirm this trend for the items presenting participants with the definition of hypothesis test and *p*-value. The concept of significance level seems again to provoke more confusion among students, with different (and greater variation in) answers and a correlation between confidence and presence of misconceptions that is close to zero.

4. Discussion and Suggestions for Further Research

This study aimed at investigating students' confidence in typical misconceptions about hypothesis tests, and more specifically about the definitions of hypothesis test, *p*-value, and significance level. We first provided empirical evidence about the presence and type of the most common misconceptions and next we explored their relation with students' confidence. Here we start the discussion by comparing our results on the appearance of misconceptions with previous research. Next we discuss more deeply the results on students' confidence.

When contrasted with earlier studies, our results confirm that there is still a substantial number of university students of introductory statistics courses holding misconceptions about the concepts of hypothesis test, *p*-value, and significance level. However, some differences with previous findings could be observed for the first two concepts. In our study, slightly lower percentages of students seemed to hold certain misconceptions as compared to earlier research. For example, concerning the first item, the percentage of participants considering the hypothesis test as a mathematical proof (*H1*) in Vallecillos (2000) was 43% (*n*=436), much higher than ours (20% and 1%, see Table 2). When we restrict her sample to those students taking a curriculum equivalent to that of our participants (in her sample *n*=167; Mathematics: 18.5%, Medicine 36.5%, and Business studies 45%) we see that the percentage decreases to 28%, which is still somewhat higher than ours. Also higher percentages of the psychology students from Haller and Krauss (2002) agreed with the statements "You have absolutely disproved the null hypothesis" (34%, *n*=44) and "You

have absolutely proven your experimental hypothesis" (20%) which could both be seen as *H1*. A higher percentage was also observed in recent study by delMas et al. (2007) where 27% of their participants (*n*=716, 32% in the post-test) believed that "Rejecting the null hypothesis means that the null hypothesis is definitely false" (*H1*). In the case of Falk and Greenbaum (1995), on the other hand, participants commit *H1* much less than in our study, with just 4% (*n*=53) agreeing with the statements "We proved that H_0 is not true" or "We proved that H_a is true". Nevertheless, a big difference can be found when looking at the popularity of *H2*, since an immense 83% selected "We showed that H_0 is improbable" or "We showed that H_a is probable", as opposed to our 19% and 3% (see Table 2). With regard to the item about the definition of *p*-value, a similar phenomenon can be perceived. For example, Haller and Krauss (2002) found higher percentages of participants showing the misconception representing *p4* by agreeing with the following statement: "If you decide to reject the null hypothesis, [the *p*-value is] the probability that you are making the wrong decision". They recorded 68% (*n*=44) of their students selecting this option, as compared to 16% in our sample (see Table 3). In the case of delMas et al. (2007), a great percentage of participants were unable to recognize a correct interpretation of a *p*-value, even after taking the course (more than 53% pretest, 45% posttest, *n*=712). In our case it was 48% (100% - 52% who selected the correct answer, see Table 3) the percentage of students not able to identify and pick the appropriate definition of this concept.

The observed differences with previous research point all in the same direction; namely, that the percentage of students committing misconceptions in our study about hypothesis tests and *p*-values is, in general, slightly lower than in the other studies considered. However, noteworthy conclusions are difficult to extract, and a detailed reading of the different publications reveals important methodological differences between our study and previous findings that make the dissimilarities hard to interpret. For example, in the study by Vallecillos (2000), instead of selecting a correct option out of a list, participants had to choose "true" or "false" for the statement "A statistical hypothesis test, when properly performed, establishes the truth of one of the two null or alternative hypotheses". In the case of Haller and Krauss (2002) not only were the items constructed in "true/false" style but also the sample of participants was quite different from ours, including only members of psychology departments (methodology instructors, scientific psychologists and psychology students; percentages given above refer only to students' responses). That is again the case for Falk and Greenbaum (1995), who, although using a type of question much more similar to ours with multiple statements, gathered data only from psychology students. In addition, the small difference between our results and those of previous research could be due not only to the style of the items or the participants' major field of study, but also to the evolution in statistics education as mentioned in the Introduction in the last (more than) 10 years that separate most of those research studies from ours.

This evidence could be taken as starting point for further research that might focus on confirming this evolution and evaluating possible causes (reform in statistics education, more statistically literate society, regional differences, etc) by means of administering the same instruments that are being used by other researchers (e.g., CAOS test, delMas et al. 2007), so that comparisons and longitudinal studies are meaningful.

Regarding our analysis of participants' confidence and its relation with the appearance of misconceptions, we did find a significant difference in confidence between those students who provided a correct answer and those who selected at least one misconception, for the first two target items (definitions of hypothesis test and *p*-value). Specifically, students who answered correctly were more confident on average about their answer than students selecting incorrect statements, which is a desirable situation (Brophy 2004). This finding further stresses the importance of designing classroom activities that would enhance self-efficacy for college students (Choi 2005), not because it proves any causal relationship but because it agrees with those theoretical considerations from previous research (Brophy 2004; Choi 2005). In fact, Brophy (2004) gives advice about how to stimulate students' confidence and take the motivational concern into account when planning curriculum, instruction, and assessment. However, we have to be careful when promoting higher levels of confidence because, as pointed out by Liddell and Davidson (2004), "Confidence may also facilitate the learning process. […] However, there is a ceiling on the amount of confidence that is beneficial. […] Educators must ensure that they are not promoting increases in confidence without concurrent increases in reflection" (Liddell and Davidson 2004, p.55). Therefore, since large efficacy misjudgments, in either direction, could have detrimental consequences (Multon and Brouwn 1991), the efforts should be focused on improving students' accuracy of self-efficacy (without damaging optimism), as suggested by Pajares (1996).

It could thus be argued that confronting students with their misconceptions and with the fact that some correct ideas might lead to misconceptions in certain situations could be a good strategy, in this case to increase their confidence in the reinforced concepts while promoting reflection about their statistical beliefs. Rossman and Chance (2004) also defended the idea that the best way to help students learn statistics is to quickly detect and address such misconceptions, based on principles such as "[…] learning occurs through struggling and wrestling with ideas" or "[…] learning is enhanced by having students become aware of and confront their misconceptions" (Rossman and Chance 2004, p. 2). They designed what they called "What Went Wrong?" exercises that presented students with sample responses (some incorrect) and asked them to identify and correct the errors. The idea behind these exercises was that students would improve their ability to question an answer before proceeding. Other innovative materials that have been adapted and implemented during the past few years for introductory statistics courses can be found in the web platform of the NSF-funded AIMS project (Adapting and Implementing Innovative Material in Statistics, http://www.tc.umn.edu/~aims/, see Garfield and Ben-Zvi 2008). These materials integrate not only discussions but also collaborative activities around relevant statistical topics, offering the opportunity to confront students with well-known misconceptions. This line of research also offers possibilities for further investigating misconceptions and how to address them.

An interesting issue that seems to arise from our results is that the concept of the significance level appears to be more confusing for the students than the definitions of the hypothesis test and *p*-value. Students did not only provide a completely correct answer less often for this item, but also their responses were more spread among options and they seemed less confident about their choices as compared to the other two questions. The significance level is a fundamental underlying concept in the testing of statistical hypothesis, and it is very closely related to the definition and interpretation of *p*-values, which, after all, are normally taken and reported as "the result" of hypothesis tests. The particular difficulties that our participants experienced when dealing with a question about the significance level concept might be related to the hybridism of Fisher's and Neyman-Pearson's approaches in today's statistical practice (Batanero 2000; Chow 1996; Falk and Greenbaum 1995; Vallecillos 2000). Fisher's use of the *p*-value as a measure of the strength of evidence against the null hypothesis that has become a routine in interpreting the results of hypothesis tests is applied together with the Neyman-Pearson's focus on decision. Also Neyman-Pearson's a priori choice of the significance level is widely used; as well as their *type I* and *type II error* terminology. As Borovcnik and Peard (1996) indicated, the different axiomatic theories do not cover the ideas of the other positions very well and the tension between them is the culprit of much confusion around the use of hypothesis tests. Hence, in further research it could be explored whether more stress on the significance level instead of on the interpretation of the *p*-value, that has received more emphasis by statistical instructors in the last years, could help students better understand the logic behind the process of testing statistical hypotheses. In addition, other types of items such as open-answer questions and/or qualitative studies would be useful to further explore this phenomenon, since it could be argued that students were confused by the similarities between our second (about the *p*-value) and third item (about the significance level).

If it is the case that the problem with the third item was the higher difficulty of the concept, and not the style or the wording of the question, another interesting line of research would be to attempt to identify external sources of self-efficacy determination other than correctness, such as difficulty of the task, other task's characteristics, stage of learning, etc (Multon 1991; Pajares 1996).

In summary, three lines of research have been suggested here. First, performing comparative and longitudinal studies about the evolution of misconceptions. Second, developing pedagogical instruments to confront students with their misconceptions and to help them adjust their self-efficacy judgments. And third, exploring factors that could better explain the relationship between confidence and accuracy for the definition of the significance level, as well as evaluating the effectiveness of more stress on this concept in the introductory statistics course. In addition, a main research interest that derives from our study is that of further investigating the relation between statistical conceptions and students' specifically related confidence. For instance, developing quantitative investigations that extend the one presented here, and confront students not only with particularly difficult or complicated concepts such as *p*-values but also with less misunderstood statistical ideas (e.g., the mean, the median, etc). These studies could shed new light on the comparison between, on the one hand, the relation between confidence and common misconceptions, and, on the other hand, the relation between confidence and less controversial statistical concepts.

Appendix

Translation of the original Spanish target items:

1. Select the definition of "hypothesis test/contrast" that you consider most correct:

   __ a. Proof of the truth or falseness of the null hypothesis

   __ b. Proof of the truth or falseness of the alternative hypothesis

   __ c. Proof of the probability or improbability of the null hypothesis

   __ d. Proof of the probability or improbability of the alternative hypothesis

   __ e. Assessment of the evidence in the data in favor of or against the null hypothesis

   __ f. Assessment of the evidence in the data in favor of or against the alternative hypothesis

   How confident are you that your answer is correct?

	0	1	2	3	4	5	6	7	8	9	10	
No confidence												100% confident

2. Researchers have registered the values of a quantitative value for two groups of individuals and carried out a test of the hypothesis that there is no difference between the groups concerning that variable. The hypothesis test results in a *p*-value of 0.01. Select the conclusion that you consider most correct:

   __ a. The probability of the null hypothesis being true is 0.01

   __ b. The probability of the null hypothesis being true, given the collected (or more extreme) data is 0.01

   __ c. The probability of obtaining the same (or more extreme) data, assuming the null hypothesis is true, is 0.01

   __ d. The probability of obtaining the same (or more extreme) data is 0.01

   __ e. The probability of committing a mistake if the null hypothesis is rejected is 0.01

   __ f. The difference between the two groups is big

   __ g. The difference between the two groups is small

How confident are you that your answer is correct?

	0	1	2	3	4	5	6	7	8	9	10	
No confidence												100% confident

3. The results of a hypothesis test are statistically significant for a significance level of a = 0.05. What does this mean? Select the conclusion that you consider most correct:

__ a. It has been proven that the null hypothesis is improbable

__ b. It has been proven that the null hypothesis is false

__ c. The probability of rejecting the null hypothesis is 95%

__ d. The probability of rejecting the null hypothesis, assuming the null hypothesis is true, is 5%

__ e. The probability that the null hypothesis is true, assuming it is rejected, is 5%

__ f. The probability that the null hypothesis is true is 5%

How confident are you that your answer is correct?

	0	1	2	3	4	5	6	7	8	9	10	
No confidence												100% confident

Acknowledgments

This study is part of, and was funded by, research GOA 3H040514 from the Katholieke Universiteit Leuven (Belgium). The authors would also like to thank the student Virginie März for her help with the analyses of the empirical data, and all teachers and students of the Complutense University (Madrid, Spain) who voluntarily participated in the study.

References

Allen, K., Reed Rhoads, T., and Terry, R. (2006). Work in progress: Assessing student confidence of introductory statistics concepts. In *Frontiers in Education Conference, 36th Annual.*

Bandura, A. (1977). Self-efficacy. Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical thinking and learning (An international journal), 2*(1 and 2), 75-97.

Ben-Zvi, D. and Garfield, J. (2004). *The challenge of developing statistical literacy, reasoning and thinking*. The Netherlands: Kluwer Academic Publishers.

Borovcnik, M. and Peard, R. (1996). Probability. In A.J.Bishop (Ed.), *International Handbook of Mathematics Education* (pp. 239-287). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, 10*(3), 252-268.

Brophy, J. (2004). *Motivating Students to Learn* (2nd ed.). Mahwah (N.J.): Erlbaum.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., and Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98-113.

Chance, B., delMas, R. C., and Garfield, J. (2004). Reasoning About Sampling Distributions. In D.Ben-Zvi and J. Garfield (eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). The Netherlands: Kluwer Academic Publishers.

Choi, N. (2005). Self-efficacy and self-concept as predictors of college students' academic performance. *Psychology in the Schools, 42*(2), 197-205.

Chow, S. L. (1996). *Statistical significance* . London: SAGE Publications Ltd.

Cohen, S., Smith, G., Chechile, R. A., Burns, G., and Tsai, F. (1996). Identifying impediments to learning probability and statistics from an assessment of instructional software. *Journal of Educational and Behavioral Statistics, 21*(1), 35-54.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2), 23-32.

delMas, R. C. (2001). Sampling SIM (Version 5) [Computer software].

delMas, R. C., Garfield, J., Ooms, A., and Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28-58.

Derry, S. J., Levin, J. R., Osana, H. P., Jones, M. S., and Peterson, M. (2000). Fostering students' statistical and scientific thinking: Lessons learned from an innovative college course. *American Educational Research Journal, 37*(3), 747-773.

Dutton, J. and Dutton, M. (2005). Characteristics and performance of students in an online section of business statistics. *Journal of Statistics Education, 13*(3).

Eccles, J. S. and Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109-132.

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

Falk, R. and Greenbaum, C. W. (1995). Significance tests die hard. *Theory and Psychology, 5*(1), 75-98.

Finney, S. J. and Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology, 28*, 161-186.

Finzer, B. (2005). Fathom (Version 2) [Computer software]. Emeryville, CA: Key Curriculum Press.

Garfield, J. (1993). Teaching statistics using small-group cooperative learning. *Journal of Statistics Education, 1*(1).

Garfield, J. (2001). *Evaluating the impact of educational reform in statistics: A survey of introductory statistics courses.* Final Report for NSF Grant REC-9732404.

Garfield, J. and Ben-Zvi, D. (2008). Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice. Emeryville, CA: Springer.

Giraud, G. (1997). Cooperative learning and statistics instruction. *Journal of Statistics Education, 5*(3).

Gliner, J. A., Leech, N. L., and Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? *The Journal of Experimental Education, 71*(1), 83-92.

González, G. M. and Birch, M. A. (2000). Evaluating the instructional efficacy of computer-mediated interactive multimedia: Comparing three elementary statistics tutorial modules. *Journal of Educational Computing Research, 22*(4), 411-436.

Gordon, H. R. D. (2001). AVERA members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Education Research, 26*(2).

Guzzetti, B. J., Snyder, T. E., Glass, G. V., and Gamas, W. S. (1993). Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly, 28*(2), 116-159.

Haller, H. and Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1), 1-20.

Innabi, H. (1999). Students' judgment of the validity of societal statistical generalization. In Alan Rogerson (ed.), *Proceedings of the International Conference on Mathematics Education into the 21st Century: Societal Challenges, Issues and Approaches*. Cairo.

Keeler, C. M. and Steinhorst, R. K. (1995). Using small groups to promote active learning in the introductory statistics course: A report from the field. *Journal of Statistics Education, 3*(2).

Kirk, R. E. (2001). Promoting good statistical practices: some suggestions. *Educational and Psychological Measurement, 61*(2), 213-218.

Lecoutre, M.-P., Poitevineau, J., and Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *International Journal of Psychology, 38*(1), 37-45.

Liddell, M. J. and Davidson, S. K. (2004). Student attitudes and their academic performance: Is there any relationship? *Medical Teacher, 26*(1), 52-56.

Mittag, K. C. and Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*, 1420.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International statistical review, 65*, 123-165.

Multon, K. D. and Brouwn, S. D. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology, 38*(1), 30-38.

National Council of Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: NCTM.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543-578.

Rossman, A. J. and Chance, B. (2004). Anticipating and addressing student misconceptions. Presented at ARTIST Roundtable Conference, August, 2004.

Schau, C., Dauphinee, T. L., Del Vecchio, A., and Stevens, J. (1999). Survey of attitudes toward statistics (SATS) [Online: http://www.unm.edu/~cschau/downloadsats.pdf].

Scheines, R., Leinhardt, G., Smith, J., and Cho, K. (2005). Replacing lecture with web-based course materials. *Journal of Educational Computing Research, 32*(1), 1-26.

Schuyten, G., Dekeyser, H., and Goeminne, K. (1999). Towards an electronic independent learning environment for statistics in higher education. *Education and Information Technologies, 4*(4), 409-424.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F.K.Lester Jr. (ed.). Second handbook of research on mathematics teaching and learning (pp. 957-1009). Greenwich, CT: Information Age Publishing, Inc. and Charlotte, NC: NCTM.

Smith III, J. P., diSessa, A. A., and Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115-163.

Stajkovic, A. D. and Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin, 124*(2), 240-261.

Vallecillos, A. (2000). Understanding of the logic of hypothesis testing amongst university students. *Journal for Didactics of Mathematics, 21*, 101-123.

Williams, A. M. (1998). Students' understanding of the significance level concept. In L. Pereira-Mendoza, L. S. Kea, T. W. Kee, and W. Wong (eds.), *Proceedings of the 5th International Conference on Teaching Statistics* (pp. 743-749). Voorburg, The Netherlands.

Ana Elisa Castro Sotos
Vesaliusstraat 2
3000 Leuven
Belgium
e-mail: anaelisa.castrosotos@ped.kuleuven.be
Phone: +3216326265
Fax: +3216325934

Stijn Vanhoof
Vesaliusstraat 2
3000 Leuven
Belgium
e-mail: Stijn.Vanhoof@ped.kuleuven.be

Wim Van den Noortgate
Vesaliusstraat 2
3000 Leuven
Belgium
e-mail: Wim.VandenNoortgate@kuleuven-kortrijk.be

Patrick Onghena
Vesaliusstraat 2
3000 Leuven
Belgium
e-mail: Patrick.Onghena@ped.kuleuven.be