Available online at www.sciencedirect.com
ScienceDirect
Educational Research Review 2 (2007) 98–113
Educational Research Review
www.elsevier.com/locate/EDUREV

# Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education

Ana Elisa Castro Sotos *, Stijn Vanhoof, Wim Van den Noortgate, Patrick Onghena

*Centre for Methodology of Educational Research, Katholieke Universiteit Leuven, Vesaliusstraat 2, 3000 Leuven, Belgium*

Received 19 December 2006; received in revised form 13 April 2007; accepted 27 April 2007

**Abstract**

A solid understanding of *inferential statistics* is of major importance for designing and interpreting empirical results in any scientific discipline. However, students are prone to many misconceptions regarding this topic. This article structurally summarizes and describes these misconceptions by presenting a systematic review of publications that provide empirical evidence of them. This group of publications was found to be dispersed over a wide range of specialized journals and proceedings, and the methodology used in the empirical studies was very diverse. Three research needs rise from this review: (1) further empirical studies that identify the sources and possible solutions for misconceptions in order to complement the abundant theoretical and statistical discussion about them; (2) new insights into effective research designs and methodologies to perform this type of research; and (3) structured and systematic summaries of findings like the one presented here, concerning misconceptions in other areas of statistics, that might be of interest both for educational researchers and teachers of statistics.


*Keywords:* Statistical inference; Misconceptions; Students; Research review

## 1. Introduction

Since the early 1970s, there has been an increasing interest in research about people's understanding and performance in probability and statistics. Researchers especially focused on misconceptions and heuristics regarding probability, chance, and randomness (e.g., Kahneman, Slovic, & Tversky, 1982; Konold, 1989, 1991; Nisbett & Ross, 1980; Shaughnessy, 2003). This article reports on a systematic review of the available empirical evidence of students' misconceptions in statistical inference. It constitutes a starting point for educational researchers interested in the relation between misconceptions and the conceptual change theory (e.g., Finch & Cumming, 1998; Guzzetti, Snyder, Glass, & Gamas, 1993; Smith, diSessa, & Roschelle, 1993), or a helpful tool for teachers of statistical inference to become aware of the most common misconceptions that their students may hold (e.g., Batanero, Godino, Vallecillos, & Holmes, 1994; Brewer, 1985; Haller & Krauss, 2002).

In educational research, the term *misconception* is used to refer to several concepts. On the one hand, authors often consider a broad definition of the word, using it to label different concepts such as *preconception*, *misunderstanding*, *misuse*, or *misinterpretation* interchangeably (Smith et al., 1993). On the other hand, misconceptions are sometimes defined in a more restrictive way, as misunderstandings generated during instruction, emphasizing a distinction with

* Corresponding author. Tel.: +32 16 32 62 65; fax: +32 16 32 59 34.
*E-mail address:* anaelisa.castrosotos@ped.kuleuven.be (A.E. Castro Sotos).

1747-938X/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.edurev.2007.04.001

alternative conceptions resulting from ordinary life and experience (Guzzetti et al., 1993). In this manuscript, a refinement of the first definition is applied, and the term refers to any sort of fallacies, misunderstandings, misuses, or misinterpretations of concepts, provided that they result in a documented systematic pattern of error (Cohen, Smith, Chechile, Burns, & Tsai, 1996).

The interest on *statistical inference* arises from three realities. First, this is a topic of main relevance for the development of research in all empirical sciences in general and psychology and education in particular (Belia, Fidler, Williams, & Cumming, 2005; Krauss & Wassner, 2002). Second, inference receives special attention in statistical courses from almost all scientific areas, where hypotheses tests and confidence intervals are taught to students as *the* methods for evaluating scientific hypotheses (Aberson, Berger, Healy, & Romero, 2003; APA, 2001). Finally, inferential ideas seem to be especially sensitive to be misunderstood and students are often prone to fall into deep misconceptions (Daniel, 1998; Kirk, 2001) because they require students to understand and connect many abstract concepts such as sampling distribution and significance level.

After presenting our methodology of search for this review (Section 2), we provide an overview of the misconceptions mentioned and exemplified in the literature, and describe to what extent and under which conditions they occur, discussing the methodology of the presented group of studies (Section 3). Finally, we conclude with some suggestions for further research (Section 4).

## 2. Method

We performed a thorough literature exploration in order to bring together publications that report on studies providing *empirical* evidence of *university students*' misconceptions that have been published during the last 15 years (from 1990 to the beginning of 2006). Therefore, studies based on personal experience and anecdotes only or publications oriented to other groups (such as professionals or younger students) were excluded.

Other publications that did not match our inclusion criteria, for instance purely theoretical discussions of misconceptions, will be used here to illustrate original ideas or describe these misconceptions.

We followed four different lines of search: first, as our main source, we surveyed the *Web of Science* (ISI, 2004), *PsycINFO* (APA, 2006), and *ERIC* (IES, 2006) electronic databases. We based our selection of key words[1] on the criteria described above and the main concepts of statistical inference. These concepts were chosen according to the logical structure of statistical inference, which is based on the foundational sampling theory and consists of two main inferential techniques: hypotheses tests and confidence intervals. This structure is as well reflected in the content and order of topics in most handbooks of introductory courses on statistical inference (e.g., Healey, 2005; Moore & McCabe, 2006). These handbooks approach statistical inference by firstly introducing sampling processes and related concepts and properties, focusing on the concepts of population, sample, point estimation by means of sample statistics (e.g., the sample mean), sample and sampling distributions, and practical approximations (e.g., central limit theorem). Next, students learn how to construct and understand confidence intervals for the estimation of different parameters as well as to perform hypotheses tests. We will use this structure to present the results of our review in this manuscript (see Section 3).

Second, after searching in electronic databases, we complemented our list of references by in depth scrutinizing the main forums for research on statistics education:

- Journal of Statistics Education, all available articles: from Volume 1 (July 1993) to Volume 13 (November 2005).
- Statistics Education Research Journal, all available articles: from Volume 1 (May 2002) to Volume 4 (November 2005).
- Proceedings from the 5th (1998) and 6th (2002) International Conferences on Teaching Statistics (ICOTS).

Next, we started a third line of search, tracking down all references cited in the results of the previous two searches.

[1] The following combination of keywords was used: (*Misconception* OR *error* OR *misuse* OR *misinterpretation* OR *misunderstanding* OR *fallacy*) AND (*inference* OR *sampling* OR *normal* OR *confidence* OR *interval* OR *test* OR *level* OR *p-value*) AND *student* AND (*course* OR *college* OR *university*).

Table 1
Summary of publications presented in this review

| Topic concerning the misconception | Publications |
|---|---|
| Sampling distributions | |
| The law of small numbers and sampling variability | Chance et al. (2004), delMas and Liu (2005), Finch (1998), Sedlmeier (1998) and Well et al. (1990) |
| The different distributions | Chance et al. (2004) and Lipson (2002) |
| The central limit theorem | Batanero et al. (2001, 2004) |
| Hypotheses tests | |
| Approaches to hypotheses testing | Vallecillos and Batanero (1997) |
| Definition of hypotheses | Vallecillos and Batanero (1997) |
| The conditional nature of significance levels | Haller and Krauss (2002), Vallecillos (2002), Vallecillos and Batanero (1997) and Williams (1998) |
| Interpretation of the numerical value of the *p*-value | – |
| Nature of hypotheses tests | Falk and Greenbaum (1995), Haller and Krauss (2002) and Vallecillos (1995, 1996, 2000) |
| Evaluation of statistical significance | – |
| Confidence intervals | |
| See Table 3 | Fidler (2006) |

Finally, as the fourth search, we looked for those journal articles, book chapters, and conference proceedings that cite classical literature about stochastic reasoning (e.g., Kahneman et al., 1982; Shaughnessy, 1992) or the results from the previous three searches, making use of two electronic search engines (*Web of Science* and *Google Scholar*). This last search complemented the previous three in finding manuscripts that could have been missed or that could have escaped from those searches.

We classified the resulting group of references following the structure of inferential statistics courses mentioned above, depending on the topic of each documented misconception (see Section 3). Besides, we recorded specific information for each of the studies: characteristics of participants (number, gender, statistical background, and studies), country where the study took place, and method of data collection (interview, multiple-choice test, open-answer questionnaire, etc.). Most of this information, when not available in the text, can be found in Appendix A.

## 3. Results

As our findings show, the literature on statistics education, and particularly publications providing empirical evidence of misconceptions in statistics, is sparse (see also Batanero, 2005; Ware & Chastain, 1991). The four searches defined above yielded more than 500 references that contained only 21 publications (cf. Table 1) reporting on 17 different studies (cf. Table A1) that provide evidence of misconceptions about topics related to statistical inference and that satisfied our selection criteria. Whereas some of the misconceptions were described several times, for others we found none or only one or two empirical studies.

Although most of the studies present the same type of study design (a one-group posttest only evaluation), the main characteristic of the research presented by this set of publications is the variation in the methodology they used. First, regarding the setting, some publications were carried out in the classroom environment (e.g., Batanero, Tauber, & Sánchez, 2001; Well, Pollatsek, & Boyce, 1990, Study 1), and a few under very controlled conditions (e.g., Vallecillos & Batanero, 1997; Well et al., 1990, Study 3) whereas other publications do not clarify in which conditions the study took place (e.g., Fidler, 2006). Moreover, a course is sometimes (e.g., Lipson, 2002), but not always (e.g., Haller & Krauss, 2002) included as a treatment stage of the study. Third, with regard to the data gathering methodology, interviews (e.g., delMas & Liu, 2005), and mixture of multiple-choice and open-answer items can be found (e.g., Vallecillos, 1995, 1996, 2000, 2002). The items are mostly presented in paper-and-pencil format (e.g., Vallecillos, 1996) and seldom by means of a computer device (e.g., Sedlmeier, 1998). Finally, there are differences in the existence of a pre-test (e.g., the SRA[2] in Vallecillos & Batanero, 1997 versus none in Falk & Greenbaum, 1995).

[2] Statistical reasoning assessment, see Garfield (2003).

Another factor making it difficult to compare results of these studies is the differences in sample size, which range from very small (e.g., delMas & Liu, 2005; Vallecillos & Batanero, 1997) to large numbers of participants (e.g., Vallecillos, 1995, 1996, 2000).

In the following, we describe the misconceptions found in this literature research. Although a strict classification would be somewhat artificial because many misconceptions overlap with each other, for the sake of clarity, they are expounded one by one, according to the logical structure for inferential statistics courses mentioned above. First, a detailed description of each misconception is illustrated with the help of all references regarding that misconception that were found, not only those satisfying the criteria in Section 2. Next, the selected publications detected by our searches are enumerated and commented more in detail, highlighting the most striking empirical results that they provide.

### *3.1. Sampling distributions*

Sampling distributions are central to statistical inference. They are used to test hypotheses and construct confidence intervals. The key idea in inferential reasoning is that a sample provides some, but not complete, information about the population from which it is drawn. Understanding this fundamental property of sampling processes implies understanding a balance between two concepts. First, *sample representativeness*, meaning that when the process of selecting the sample has been performed properly, the sample will often have characteristics similar to those of the population; and, second, *variability*, implying that not all the samples can resemble the population in the same way and to the same extent every time (Batanero et al., 1994).

Many statistics educators have stressed that the sampling distribution is a core idea in the understanding of statistical inference. Yet, despite its critical role, experience and research have shown that this concept is generally poorly understood (Chance, delMas, & Garfield, 2004; Lipson, 2002; Shaughnessy, 1992; Tversky & Kahneman, 1971). Although many students are able to apply the necessary notions to deal with sampling processes in isolation, they often lack the ability to integrate the different ideas and properly use concepts in inferential reasoning (Batanero, 2005). Many authors claim that the explanation for this lack is the presence of misconceptions pertaining to such processes, which are not sufficiently known by teachers (Batanero et al., 1994). The following compilation of empirical evidence might be, therefore, of much use for teachers of statistical inference.

#### *3.1.1. Misconceptions concerning the law of small numbers and sampling variability*

The sample mean is a key descriptive statistic for inferential analyses. However, many misconceptions regarding sampling processes concern the sample mean, more specifically, the properties of its sampling distribution. The most important of such properties is the so-called *law of large numbers*. This law states that, for any population with finite mean, the population mean is the limit of the sample mean as the sample size increases with new observations drawn at random from the population, therefore implying that the variability of the sample mean tends to zero as the sample size increases. This final implication for the variability of the sample mean has been proven to be misunderstood by many students, who neglect the effect of sample size on the variance of the sample mean (Chance et al., 2004).

Our search criteria identified five publications documenting empirical evidence about misconceptions regarding the idea behind the law of large numbers, more specifically: Well et al. (1990), Finch (1998), Sedlmeier (1998), Chance et al. (2004; see next section), and delMas and Liu (2005).

Concerning the variability of the sample mean, the university psychology students surveyed by Well et al. (1990) seemed to understand that the means of larger samples are more likely to resemble the population mean. Moreover, some of the participants (also students of psychology) in the study by Finch (1998) considered that variance can decrease in some sense or that reliability gains from large samples. However, they did not seem to understand its implication for the variability of the sample mean, even when they seemed to have a clear understanding of the distinction between sample and sampling distributions and had observed evidence of the phenomenon (Well et al., 1990).

These difficulties might have their origin in a misunderstanding of the law of large numbers. According to the *representativeness* heuristic (Tversky & Kahneman, 1971), people confuse the sample and the population distributions, believing that any sample must be very similar to the population, regardless of its size, and therefore extrapolating the law of large numbers to small samples. This misconception is known as the *belief in the law of small numbers*, a term coined by Tversky and Kahneman (1971). The believer in the law of small numbers underestimates the size of confidence intervals, overestimates the significance in tests of hypothesis, and is over-confident in obtaining the same

results in future replications of the experiment. A student holding such belief will not see the sample size or the sample bias as factors that affect the validity of inferential conclusions or generalizations to the population (Innabi, 1999).

Gigerenzer and Sedlmeier (Gigerenzer, 1996; Sedlmeier, 1998; Sedlmeier & Gigerenzer, 2000) argue that the occurrence of this misconception depends on the type of task. These authors make a distinction between *frequency distribution tasks* (e.g., problems in Bar-Hillel, 1979), which deal with the frequency of an event, and *sampling distribution tasks* (e.g., in Kahneman & Tversky, 1972), which concern the frequency with which particular events' sample means or proportions fall into specified categories. They suggest that the intuition that larger samples usually lead to more exact estimates of the population mean or proportion (Finch, 1998; Well et al., 1990) helps in the solution of the frequency rather than the sampling distribution tasks because the law of large numbers can be applied to frequency distribution tasks directly, whereas sampling distribution tasks require repeated application of it, as well as a judgment about the variances of the distributions. In addition, they provide empirical evidence of this phenomenon in the three studies described in Sedlmeier (1998).

The misconceptions described above are therefore related to, and might arise from, a profound lack of insight in the idea of variability in random events that has a direct impact on the understanding of sampling processes and hence on the relevant properties of the sample mean distribution such as the law of large numbers. Before being able to understand the concept and features of sampling distributions, students should be able to develop a good understanding of the idea of variability (delMas & Liu, 2005). For that reason, they are expected to arrive at university introductory courses with a comfortable domain of this concept. However, several studies like Innabi (1999) or Watson and Moritz (2000) demonstrate that this is not yet the case and that there is still a lot of work to do in making students understand variability in pre-university education. It seems that even though natural cognitive development might be favorable to improving the understanding of sampling variability (delMas & Liu, 2005; Watson & Moritz, 2000), statistics courses still fail in helping and supporting students to construct understanding of the basic and core concepts of probability and statistics such as the idea of variability.

### *3.1.2. Misconceptions concerning the different distributions*

As a consequence of the representativeness misconception described above, which provokes the confusion of a specific *sample* distribution and the corresponding *population* distribution, two severe misconceptions can arise. First, if students believe that the sampling distribution of a statistic should have the same shape and properties as the population distribution, they will confuse the *population* and the *sampling* distributions (Chance et al., 2004). Second, as a result of the combination of the representativeness misconception and the confusion of population and sampling distributions, a student might not be able to detect the difference between the distribution of a *sample* and the *sampling* distribution of a statistic. We found two studies related to these misconceptions: Lipson (2002) and Chance et al. (2004).

Lipson (2002) attempted to improve the development of the concept of sampling distribution and facilitate the formation of links between the sampling distribution and statistical inference. She exposed a group of weak-mathematical students to an instructional treatment with two computer sampling packages and found that, although 10 out of the 23 participants correctly linked the sampling distribution as the distribution of a sample statistic, 5 showed evidence of one of the other two misconceptions described above, incorrectly designating the distribution of the sample as the sampling distribution. Moreover, in the same study, only 7 and 3 out of the 23 students explicitly linked the sampling distribution to the determination of the concepts of $p$-value and confidence interval, respectively. Lipson concluded that it seems possible that the extensive use of the sampling software was helpful in elucidating some important concepts of sampling distribution in each of the specific contexts in which it was applied for some students. However, these software packages have no specific role in illustrating the concepts and links, which together form a schema for the generalized sampling distribution.

In turn, Chance et al. (2004) performed a series of five studies to document students' learning of sampling distributions. More specifically, their third study was centered on a conceptual analysis of students' prior knowledge and misconceptions. The results of their post-treatment assessment tasks showed that many students still exhibited misconceptions concerning sampling distributions. In particular, they believed that a sampling distribution should look like the population (and even more as the sample size increases) showing that they did not understand that a sampling distribution is a distribution of sample statistics and confusing it with the population distribution. One example of a statement wrongly selected by the students that showed this misconception is "As the sample size increases, the sampling distribution of means looks more like the population, has the same mean as the population, and has a standard deviation that is similar to the population" (Chance et al., 2004).

*3.1.3. Misconceptions concerning the central limit theorem*

Not only does the law of large numbers explain the sampling behavior of the sample mean and its variability, but also the *central limit theorem* provides a useful and often used approximation for the sampling distribution of the sample mean. This theorem states that, for sufficiently large sample sizes, the sampling distribution of the sample mean can be approximated by a Normal distribution, which facilitates inferential calculations. Studies satisfying our searching criteria that provide empirical evidence of misconceptions with regard to the Normal distribution, and therefore in relation to this theorem, are Batanero et al. (2001) and Batanero, Tauber, and Sánchez (2004).

One of the misconceptions related to the understanding of the central limit theorem is that students wrongly extrapolate it and believe that, the larger the sample size, the closer the distribution of any statistic will approximate a Normal distribution (Bower, 2003). A related misconception is that, because they believe it always can be applied, students seem to be confused when they try to find the reason to use such an approximation and give a justification for the use of the Normal distribution, although they might be comfortable doing the formal manipulations needed to use this theorem (Wilensky, 1997).

Besides, students have shown the misconception of not properly distinguishing between the real sampling distribution, and the theoretical model of a Normally distributed population that is only used as an approximation of this sampling distribution, based on the central limit theorem that is used to test the null hypothesis in a significance test. This misconception might be related to a more general misconception about the differences between theoretical (mathematical) Normal distributions and empirical (for the actual data) almost-Normal distributions as documented in Batanero et al. (2001, 2004). They found that, for example, their first year university students did not identify the Normal model with an equation (the analytical expression of the Normal density function) and approximately 81% of them ($n = 55$) selected the statement "The Normal curve is a model that is defined by empirical data" and were not aware that the Normal curve is a theoretical distribution.

In summary, although students might be able to perform all necessary manipulations and formal calculations for testing a hypothesis, it has been shown that many of them hold deep misconceptions related to sampling distributions. The main ones being those concerning the law of large numbers, the confusion of these distributions with the sample or the population distributions, and the central limit theorem. These misconceptions have a direct impact on learning inferential statistics because of the interconnection of the concepts and methods and the relevance of their understanding for an appropriate interpretation of inferential results and conclusions. For example, a student that confounds the sampling distribution of a statistic and the population distribution of the variable under study will believe that "something is wrong in the process" when, for instance, a graphical representation of these distributions shows a great difference between them. Another example would be a student who uses the central limit theorem under the wrong conditions (because of the belief in its universal validity) and reaches wrong conclusions for the population of study. In a similar way, a student confounding the sample and population distributions is prone to overlook the sample bias as a factor that affect the validity of inferential conclusions.

*3.2. Hypotheses tests*

The main tool in inferential statistics is the hypotheses test (also called *significance* test). This technique aims to state the evidence in a sample against a previously defined (null) hypothesis, minimizing certain risks. Getting students to make sense of hypotheses tests is a very difficult goal for statistics instructors because of the persistency and deepness of the misconceptions hold by learners (Brewer, 1985; Daniel, 1998; Kirk, 2001), even after years of training (Falk, 1986; Vallecillos, 2002). The main reason for this phenomenon is that performing these tests requires students to understand and be able to relate many abstract concepts such as the concept of a sampling distribution, the significance level, null and alternative hypotheses, the $p$-value, and so on.

For more than 20 years, misconceptions and misuses regarding hypotheses tests have been discussed. An historical summary of bad practices can be found in Daniel (1998) and an overview of the controversy about the use of these tests in research in Falk (1986), Harlow, Mulaik, and Steiger (1997), Kirk (2001), McLean and Ernest (1998), and Reichardt and Gollob (1997). According to Krantz (1999), misconceptions concerning hypotheses tests are not the fault of the method, but the responsibility of those who use it. Several authors (e.g., Brewer, 1985; Gliner, Leech, & Morgan, 2002) have pointed at textbooks as main culprits of misconceptions, while others (e.g., Haller & Krauss, 2002) claimed that there is also a large number of statistics instructors that share the misconceptions of their students, and have an even larger influence on fostering the misconceptions than the textbooks have. In fact, Gordon (2001), Lecoutre, Poitevineau,

and Lecoutre (2003), and Mittag and Thompson (2000) convincingly showed that even statisticians are not immune to misconceptions of hypotheses tests.

The following compilation of publications provides an overview of the main misconceptions held by university students (and apparently also by professional researchers) with regard to hypotheses tests. In the next sections it will be highlighted when, for a misconception, no empirically based publications were found by our searches and its description is based on purely theoretical studies or on results from other type of participants different from university students (e.g., Misconception 5 from exploration of textbooks in Gliner et al., 2002).

*3.2.1. Misconception concerning the different approaches to hypotheses testing*

As shown in Cohen (1990) and Chow (1996), students' misconceptions about hypotheses tests can be classified and analyzed from several points of view, but, in general, most misconceptions about learning and using these tests have been attributed in the literature to two aspects: the philosophy behind the test, and the interpretation of concepts and results. Regarding the philosophy of the test, hybridism between Fisher's and Neyman–Pearson's approaches is mostly present in statistical practice (Batanero, 2000; Chow, 1996; Falk & Greenbaum, 1995; Vallecillos, 2000). The systematic Fisher's comparison of the $p$-value with the level of significance that has become a routine behavior 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 mixture is the culprit of much confusion around the use of hypotheses tests (see also Gigerenzer, 1993).

Vallecillos and Batanero (1997) is the only publication found by our searching criteria aiming to empirically address this (and the following) misconception. They found students' conception far from the logic normally used in teaching that considers hypotheses testing as a decision process in order to accept or reject a hypothesis, according to the Neyman–Pearson's theory. Some of their students considered the hypotheses as alternatives in a decision problem under uncertainty, whereas the majority of them did not acknowledge the parallelism between the hypotheses testing and the decision process.

*3.2.2. Misconceptions concerning the definition of the hypotheses*

With regard to misconceptions about concepts and results, there are elements in every stage of hypothesis testing that students misunderstand. First, in Vallecillos and Batanero (1996) it is remarked that the confusion between the null and the alternative hypotheses turns out to be a serious misconception that obstructs the understanding of the testing process and specially the correct interpretation of its results. In fact, the election of inadequate hypotheses determines the results of the complete process and in Vallecillos (1999) it is assured that the first step of defining the null and alternative hypotheses presents great comprehension problems for students, who are unable to identify the most appropriate statement for each case.

In Vallecillos and Batanero (1997) evidence is provided of students' incorrect choice of hypothesis. Although participants in their study agree on the theoretical idea of stating a null hypothesis with the intention of finding evidence against it, they are not consistent when they are asked to define the null and alternative hypothesis for a specific contextualized problem, exchanging those two concepts. Moreover, another two specific misconceptions were detected in Vallecillos and Batanero (1997):

- Believing that the null hypothesis and the acceptance region "are the same thing".
- Believing that a hypothesis can refer both to a population and a sample.

*3.2.3. Misconceptions concerning the conditional nature of significance levels*

Once the hypotheses have been specified and the analytical calculations for the test (including the estimation of the sampling distribution of the test statistic) have been performed, it is time for the student to interpret the obtained results. The most complicated concepts associated to hypotheses testing results are the significance level ($\alpha$) and the $p$-value ($p$) (Haller & Krauss, 2002).

The most common misconception of the significance level and the $p$-value is that of switching the two terms in the conditional probabilities (Falk, 1986). As a consequence, $\alpha$ is considered as the probability that the null hypothesis is true once the decision to reject has been taken, and $p$ as the probability that the null hypothesis is true given the observed

data. Our searches have detected empirical evidence of this switch in Vallecillos and Batanero (1997), Williams (1998), Vallecillos (2002), and Haller and Krauss (2002).

The occurrence of this misconception is impressive. For example, 30 out of the 44 students surveyed by Haller and Krauss agreed with a statement saying "You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision" referring to the meaning of the *p*-value of the test (Haller & Krauss, 2002). In the study by Vallecillos 53% ($n = 436$) of the students selected a similar item stating "A level of significance of 5% means that, on average, 5 out of every 100 times that we reject the null hypothesis, we shall be wrong" (Vallecillos, 2002). In these studies, university students from different disciplines following statistics courses were selected (only from psychology in the case of Haller and Krauss).

Using the same items later mentioned in Vallecillos (2002) and Vallecillos and Batanero (1997) carried out a series of interviews with seven excellent students, most of which interpreted the significance level as the probability of the null hypothesis being true given that it has been rejected. In other words, falling into the misconception and switching the terms in the conditional probability defined by $\alpha$. Finally, Williams (1998) also interviewed university students about the concept of significance level and also found several quotes reflecting the switch of the conditional probabilities misconception.

The misconception of confusing a conditional and its inverse might reveal a deeper misconception inherited from probability, documented in Shaughnessy and Dick (1991), that was recently ascribed to the language most textbooks and trainers use to describe conditional probabilities, in Ancker (2006; see also Falk, 1986; Falk & Greenbaum, 1995). This author identifies the switch of the conditional (and the *ignoring*, see next section) as a main misconception when, for example, students think the conditional probability of having a disease given a positive diagnostic test to be the same as the conditional probability of a positive diagnostic test given the disease (or just the probability of having the disease, see next section). In Ancker's opinion, this misconception might be triggered by the textbook and teacher's use of the expression "A given B" for the probability $P(A|B)$ which may not be sufficient to draw attention to the existence of a new category of events given by the change in the sample space introduced by the conditional (Ancker, 2006). This author's advice is to describe conditional events as "A within B" or "A in B" which can encourage students to visualize one event as a subset of another, thus emphasizing the existence of a new category defined by the conditioning event.

Ignoring the conditional and considering the significance level and the *p*-value as single event probabilities generates another important group of misconceptions concerning the conditional nature of significance levels:

- $\alpha$ is regarded as the probability that the one of the hypotheses is true.
- $\alpha$ is regarded as the probability of making a mistake.
- $p$ is regarded as the probability that the observed event happened by chance (an interpretation that is not necessarily incorrect but is vague since it does not specify the nature of the conditional).

In explaining why these misconceptions prevail among students, statistical terminology was already blamed in Falk (1986), where it was suggested that they might be well explained by the Neyman–Pearson's way to denote $\alpha$ as the probability of a *type I error*. This expression is not conditionally phrased and could reasonably be interpreted as the conjunction of the two uncertain events (i.e., that the null hypothesis is true *and* that it is rejected). In Falk and Greenbaum (1995) it is discussed more in depth the role of verbal ambiguity in statistical misconceptions and introduced the *illusion of probabilistic proof by contradiction* (see Misconception 6) as a possible serious consequence of it.

Our searches found only one publication that empirically investigates this type of misconceptions. This is the study (Haller and Krauss, 2002) mentioned above, where 14 out of the 44 psychology students believed that a hypotheses test provides the single probability of the alternative hypothesis being true and 26 agreed that $\alpha$ is the single probability of the null hypothesis.

### *3.2.4. Misconception concerning the interpretation of the numerical value of the p-value*

The last misconception of the *p*-value is that of students considering its numeric value as an indicator of the strength of the treatment effect under test, or the distance between the groups that are being compared. This misconception is addressed by Gliner and colleagues in their study of statistical textbooks (Gliner et al., 2002). They state that outcomes with lower *p*-values are sometimes interpreted by students as having stronger treatment effects than those with higher

Table 2
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 |

$p$-values. On the other hand, the strong disagreement of AERA (American Educational Research Association) members with the statement that $p$-values directly measure study effect size is reported in Mittag and Thompson (2000).

Despite of the description of the misconception provided in Gliner et al. (2002) based on their findings regarding textbooks, no empirical studies trying to demonstrate that university students fall, in fact, into this misconception, have been found by our searches.

### *3.2.5. Misconceptions concerning the nature of hypotheses tests*

The misconceptions around the key concepts of significance level and $p$-value might have a direct consequence on the conclusions that students extract from hypotheses tests. Mainly, there are two misconceptions pertaining to the information that a statistical test provide, related to the interpretation of the $p$-value. First, considering the test as a mathematical (logical) proof (Vallecillos, 1995), and second, as a probabilistic proof (Falk & Greenbaum, 1995) of one of the hypotheses. When students consider the test as a mathematical proof, they assume that, just as any mathematical procedure, the test results ($p$-value) are deterministic. Therefore, they believe that the null hypothesis has been proved to be true or false. On the other hand, if students consider a statistical test as a probabilistic proof it means that they are falling into the so-called *illusion of probabilistic proof by contradiction* (or the *illusion of attaining improbability*). This illusion is a consequence of the similarity in the formal structures of hypotheses tests' reasoning and the mathematical proof by contradiction, which is based on the logical *modus tollens* method. As can be seen in Table 2, the misconception arises when an analogous method is applied to hypothesis test results.

The analogy to the mathematical proof by contradiction does not work for hypotheses tests because a contradiction disproves the premise from which it is drawn, but a low probability event does not make the premise from which it is drawn improbable. This property shows that there is an important difference between statistical inference and propositional logic, since they are equivalent only when probabilities are 0 or 1 (Nilsson, 1986).

Our searches detected five publications addressing these two misconceptions (mathematical and probabilistic proof): Vallecillos (1995, 1996, 2000), Haller and Krauss (2002), and Falk and Greenbaum (1995). First, Vallecillos provided evidence of the mathematical conception of hypotheses tests finding that more than 42% ($n = 436$) of the students participating in the survey selected the statement "A hypotheses test, correctly performed, establishes the truth of one of the two hypotheses, null or alternative". This percentage was specially striking for the pedagogy (55.8%) and psychology (74.3%) students (Vallecillos, 1995, 1996, 2000). In turn, in the study by Haller and Krauss (2002) a large number of students agreed with the statements that the test absolutely disproves the null hypothesis (15 out of 44) or absolutely proves the alternative (9 out of 44).

On the other hand, most (26 out of 53) of the psychology students responding to the survey by Falk and Greenbaum (1995) believed that when a test of significance is conducted and the result turns out significant for a predetermined level, the null hypothesis has been proved to be improbable, showing evidence of the probabilistic proof by contradiction misconception.

### *3.2.6. Misconception concerning the evaluation of statistical significance*

One of the major misconceptions mentioned in the literature, that students might encounter when evaluating hypotheses tests, is that of understanding the difference between statistical and practical significance. The misconceptions that the calibration and evaluation of the difference between statistical significance and practical significance entail are stressed in Gliner et al. (2002). The first one does not imply the second one, neither the other way around. Profound knowledge about the contextualization of the test and the design of the experiment (sample sizes, etc.) as well as the encountered effect sizes, are needed in order to know when a statistically significant result is also practically signifi-

cant. In turn, a practically significant result might turn out not to be statistically significant. This is the reason why it is proposed in Batanero et al. (2004) to drop the word *significant* from data analysis vocabulary and use it only in its everyday sense to describe something actually noteworthy or important.

Students might be not aware of this important reality (Kirk, 2001) and even statisticians seem to have very little idea of how the interpretation of *p*-values should depend on sample size and how important effect sizes are, extracting wrong conclusions from their analyses (Lecoutre et al., 2003). Unfortunately, despite the stress put on the relevance of this misconception by some authors like Gliner et al. (2002) or Lecoutre et al. (2003), there is no empirical evidence of it regarding university students in any of the publications detected by our searches.

In summary, although students might be able to perform all necessary manipulations and formal statistical calculations concerning hypotheses tests, they have been proved to hold deep misconceptions about the meaning, use, and main concepts of this technique.

### 3.3. Confidence intervals

Due to the damaging over-reliance on the extensively misunderstood hypotheses tests, a wider use of confidence intervals has been recommended by many authors in order to improve research communication (Cumming, Williams, & Fidler, 2004; Harlow et al., 1997). Confidence intervals are strongly advocated as the best reporting strategy (APA, 2001), as an alternative, or complement to hypotheses tests in statistical inference (Reichardt & Gollob, 1997).

However, also confidence intervals are not always properly interpreted (Fidler, 2006; Fidler, Thomason, Cumming, Finch, & Leeman, 2004) and are prone to misconceptions. According to Fidler (2006), statistical reform has been advocated to a large degree based on the compelling argument of the tendency of hypotheses tests to be misinterpreted, but without providing an alternative that was shown to be less misunderstood, or to be easier to explain and teach to students.

Few researchers have studied misconceptions concerning confidence intervals, and the ones who have done so, such as Cumming and colleagues (Belia et al., 2005; Cumming, 2006; Cumming & Maillardet, 2006; Cumming et al., 2004; Fidler et al., 2004), have mostly focused on researchers' understanding instead of university students. For example, in Cumming et al. (2004) a deep misconception was found concerning the question "What is the probability that the next replication mean will fall within the original 95% confidence interval?" An internet-based investigation of researcher' answers to this question (given a graphical bar representation of the original 95% confidence interval) showed a large proportion of participants believing that about 95% of replications means will fall within the original confidence interval. This misconception is consistent with the law of small numbers intuition of underestimating sampling variability, and therefore believing that replications will be unjustifiably close to a first result (Cumming et al., 2004). However, the probability is close to .95 only when the original mean happens to fall very close to $\mu$, and the capture rate drops as it falls further from $\mu$. In fact, the probability is only .834 (Cumming & Maillardet, 2006). In Cumming et al. (2004) they conclude that graphical representation of confidence intervals in figure bars possibly prompts the misconception in a large majority of researchers. In the similar study (internet research) by Belia et al. (2005), researchers' understanding of the graphical comparison of confidence intervals for two separated means was studied. Participants widely believed that no overlapping of two 95% confidence intervals on independent means implies a significant difference at the .05 level between the means and that overlapping of the two intervals implies that there is no significant difference. However, this rule is incorrect. Although non-overlapping of the two confidence intervals does imply a significant difference, an overlap does not necessarily imply that there is no statistically significant difference at the .05 level.

The only study that our searches have found concerning students' understanding of confidence intervals, Fidler (2006), suggests that confidence intervals help decreasing the misconception that a statistically not significant result is equivalent to not practically significant (or 'no effect'). When asked to interpret results in confidence interval format, participants in this research misinterpreted statistically non-significant results (from a low powered study with a non-trivial effect size) much less than participants who had to interpret the results in hypotheses test format (10 versus 24, out of 55).

Nevertheless, a second series of experiments with psychology and ecology students revealed to Fidler that confidence intervals themselves are prone to a new set of misconceptions. Table 3 summarizes the percentage of students ($n = 180$) choosing each description of a confidence interval from a prepared list that included several misconceptions.

Table 3
Misconceptions of confidence intervals detected by Fidler (2006)

| Description of a confidence interval | Percentage of students ($n = 180$) |
|---|---|
| Plausible values for the sample mean | 38 |
| Range of individual scores | 8 |
| Range of individual scores within one standard deviation | 11 |
| The width of a confidence interval increases with sample size | 20 |
| The width of a confidence interval is not affected by sample size | 29 |
| A 90% confidence interval is wider than a 95% confidence interval (for the same data) | 73 |

Table 4
Overview of misconceptions suggested by the reviewed studies

| | |
|---|---|
| Sampling distributions | |
| The law of small numbers and sampling variability | Neglect the effect of sample size on the variance of the sample mean<br>Belief in the law of small numbers |
| The different distributions | Confuse the population and the sampling distributions<br>Confuse the sample and the sampling distributions |
| The central limit theorem | Belief that the larger the sample size, the closer any distribution is to the Normal<br>Inability to justify the use of the theorem and the Normal<br>Confusion between the theoretical and the approximated Normal |
| Hypotheses tests | |
| Approaches to hypotheses testing | Neglect the parallelism between hypotheses test and decision process |
| Definition of hypotheses | Confusion in the definition of null and alternative hypotheses<br>Confusion of the null hypothesis and the acceptance region<br>Believing that a hypothesis can refer both to a population and a sample |
| The conditional nature of significance levels | Inverse the conditional of the $p$-value<br>Inverse the conditional of the significance level<br>Interpreting the significance level as the probability of one hypothesis<br>Interpreting the significance level as the probability of making a mistake<br>Interpreting the $p$-value as the probability that the event happened by chance |
| Interpretation of the numerical value of the $p$-value | Interpreting the numeric value of the $p$-value as strength of treatment |
| Nature of hypotheses tests | Considering the test as a mathematical proof<br>Illusion of probabilistic proof by contradiction |
| Evaluation of statistical significance | Confuse practical and statistical significance |
| Confidence intervals | |
| See Table 3 | |

## 4. Conclusion

The group of publications analyzed in this review provides empirical evidence of deep and spread students' misconceptions in statistical inference. This evidence demonstrates that, although they may be able to manipulate and carry out calculations with statistical data, students have severe misconceptions concerning the interpretation of results from inferential techniques. According to this literature, the following summary and classification of students' misconceptions in statistical inference can be suggested (Table 4).

As the results of our searches showed, the available number of publications providing empirical data is still small; therefore, research should still fill some gaps and present empirical evidence about misconceptions that have not been much documented so far. These are the following: mixing approaches to hypotheses tests, confusing the hypotheses, ignoring the conditional event in the $p$-value and significance level descriptions, interpreting the $p$-value as the strength of the effect, evaluating the meaning of statistical significance versus practical significance, and misconceptions concerning confidence intervals. Besides, this review made clear the need for more empirically based studies that shed

light on the sources of misconceptions. Students' misconceptions might have their origin in textbooks, as suggested by Brewer (1985) and Gliner et al. (2002); in teachers bringing the misconceptions into the classroom (Haller & Krauss, 2002); or in inherited intuitions as documented in many publications; some of which are cited in this review to exemplify and clarify the evolution of certain misconceptions of statistical inference (e.g., Innabi, 1999; Watson & Moritz, 2000, in the first group of misconceptions of Section 3.1, or Shaughnessy & Dick, 1991, in the third group of misconceptions of Section 3.2). In addition, empirical studies should be set up to find possible means to specifically address and help students overcoming the misconceptions within the statistics course, via observational and intervention studies with large groups of participants from different backgrounds.

A second conclusion concerns the methodological aspect of this type of research and it is that more evidence should be sought in order to find confirmation of good methodologies to perform such type of studies that detect, describe, and try to find helpful activities to defeat, students' misconceptions. Apparently, the classical means to do so do not provide enough information about the misconceptions or are not enlightening in the search for solutions, since misconceptions keep on being found. As mentioned above, in addition to continuing the existing type of research (with questionnaires, interviews, etc.), special emphasis should be put on carrying out new kinds of studies, such as design experiments, to analyze possible means to tackle the detected misconceptions. One of the approaches to explore is that of confronting students with their misconceptions, considered by some authors as the most effective strategy to achieve the accurate conceptions (Finch & Cumming, 1998; Guzzetti et al., 1993). Our recommendation for statistics educators is to include both computational and non-computational items in their in-classroom assessments, so that their students' misconceptions can be identified. They could make use, for example, of the tools created by the ARTIST[3] team, who made available on-line topic scales and the Comprehensive Assessment of Outcomes in a First Statistics course (CAOS), designed for exactly that purpose.

Third, the spread nature of these kind of publications makes it necessary to promote and consider structured and systematic summaries of evidence like the one presented here, concerning students' misconceptions in the different areas of statistics (correlation analysis, measures of variability, etc.), as the best starting point for further research about misconceptions and their relation with the conceptual change theory, and for instructional design. This type of reviews shows the remaining gaps where research can be done and helps teachers of statistics in providing practical summaries of the available findings about students' misconceptions within the specific areas of statistics.

## Appendix A

The following outline summarizes the main characteristics of participants and methodology for the 20 selected studies (Table A1).

[3] Assessment resource tools for improving statistical thinking (available on line: http://app.gen.umn.edu/artist/).

Table A1
Main characteristics of the selected studies

| Studies | Participants | Students' area of study | Statistical background | Design | Instruments |
|---|---|---|---|---|---|
| Batanero et al. (2001) | 55 | Variety | Pre-university | *1 group*: pre-test + course + evaluation | *Pre-test*: SRA; *Course*: 10 sessions of 2 h (half with Statgraphics and half in traditional classroom); *Evaluation*: open answer 20-item questionnaire + report of three tasks solved with Statgraphics |
| Batanero et al. (2004) | 117 | Variety | Pre-university | *1 group*: pre-test + course + evaluation | *Pre-test*: SRA; *Course*: 10 sessions (6 of 1.5 h and 4 of 1 h) half in PC lab and half in traditional classroom; *Evaluation*: 3 open-ended tasks to solve with Statgraphics + 21-item questionnaire |
| Chance et al. (2004) | More than 100 | Variety | None assumed | *3 similar settings*: activity with PC + assessment | *Software*: simulation assessment: posttest and final exam |
| delMas and Liu (2005) | 12 | – | Pre-university | *1 group*: interview | Interview while interaction with computer program |
| Falk and Greenbaum (1995) | 53 | Psychology | Three courses | *1 group*: questionnaire | *Questionnaire*: one multiple-choice item about the result of a test |
| Fidler (2006) | 55 (1st study); 180 (2nd) | Ecology (1st); ecology and psychology (2nd) | At least introductory university statistics | *1st: 1 group*: comparison of two versions of one question; *2nd: 1 group*: questionnaire | *Question*: interpret a test or confidence interval result (comparison); *Questionnaire*: selection of statements (true/false) |
| Finch (1998) | 20 | 3rd year psychology | – | *1 group posttest only*: 1 question + interview | *Question*: rate statements about description of research study on a four-point scale (true to false); *Interview*: explain the rating |
| Haller and Krauss (2002) | 44 | Psychology | – | *3 groups* (instructors, scientists not teaching, students) *comparison*: questionnaire | *Questionnaire*: six statements to rate as true or false about a hypotheses test's results |
| Lipson (2002) | 23 | (Part-time) | None or little | *1 group*: course + exercises | *Exercises*: concept mapping exercises solved in-class |
| Sedlmeier (1998) | 46 (1st); 22 + 40 (2nd); 31 (3rd) | Variety | – | *1st and 2nd: 2 groups comparison*: 2 (1st) or 3 (2nd) tasks; *3rd: 2 groups comparison*: interview | *1st*: two version of the tasks in a PC (click the right of three possible answers); *2nd*: idem with three tasks; *3rd*: interview about real experimentation with two versions of two tasks (the control group did not have to do one of the questions) |
| Vallecillos (1995, 1996, 2000, 2002) | 436 | Variety | One introductory university course | *1 group*: questionnaire + interview | *Questionnaire*: 20-item (true/false, multiple-choice, and open answer) |
| Vallecillos and Batanero (1997) | 7 | Medicine | Very good | *1 group*: questionnaire + interview | *Questionnaire*: three true/false items and two problems |
| Well et al. (1990) | 114 (1st); 151 (2nd a); 138 (2nd b); 120 (3rd) | Psychology | No previous university statistics | *1st: 1 group*: questionnaires in groups of 5–15 (half the groups received the items in inverse order); 2nd: 2 groups *comparison*: questionnaire (two versions); *3rd: 2 groups comparison*: controlled conditions: four problems in groups 5–15 (two versions) | *1st and 2nd*: two open-answer problems; *3rd*: four open-answer problems |
| Williams (1998) | 18 | – | – | *1 group*: interviews | *Interviews*: clinical with one concept mapping task and two textbook hypothesis test's task |

## References


Aberson, C. L., Berger, D. E., Healy, M. R., & Romero, V. L. (2003). Evaluation of an interactive tutorial for teaching hypothesis testing concepts. *Teaching of Psychology*, *30*(1), 75–78.

Ancker, J. S. (2006). The language of conditional probability. *Journal of Statistics Education*, *14*, retrieved July 28, 2006, from http://www.amstat.org/publications/jse/v14n2/ancker.html

APA. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

APA. (2006). *PsycINFO*. APA., retrieved November 27, 2006 from http://www.apa.org/psycinfo/ [on-line]

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance*, *24*, 245–257.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning (An International Journal)*, *2*(1/2), 75–97.

Batanero, C. (2005). Statistics education as a field for research and practice. In *Proceedings of the 10th international commission for mathematical instruction*. Copenhagen, Denmark: International Commission for Mathematical Instruction.

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. R., & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematics Education in Science and Technology*, *25*(4), 527–547.

Batanero, C., Tauber, L. M., & Sánchez, V. (2001). Significado y comprensión de la distribución normal en un curso introductorio de análisis de datos [Meaning and understanding of the normal distribution in an introductory data analysis course]. *Quadrante*, *10*(1), 59–92.

Batanero, C., Tauber, L. M., & Sánchez, V. (2004). Students' reasoning about the normal distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 257–276). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*, 389–396.

Borovcnik, M., & Peard, R. (1996). Probability. In A. J. Bishop (Ed.), *International handbook of mathematics education* (pp. 239–287). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bower, K. M. (2003). *Some misconceptions about the Normal distribution*. Paper presented at the Six Sigma Forum. Milwaukee, WI: American Society for Quality.

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

Chance, B., delMas, R. C., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312.

Cohen, S., Smith, G., Chechile, R. A., Burns, G., & 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.

Cumming, G. (2006). Understanding replication: Confidence intervals, *p*-values, and what's likely to happen next time. In *Proceedings of the seventh international conference on teaching statistics*. International Association for Statistical Education.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, *11*, 217–227.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*, 199–311.

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., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55–82.

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

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

Fidler, F. (2006). Should psychology abandon *p*-values and teach CIs instead? Evidence-based reforms in statistics education. In *Proceedings of the seventh international conference on teaching statistics*. International Association for Statistical Education.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119–123.

Finch, S. (1998). Explaining the law of large numbers. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the fifth international conference on teaching statistics* (pp. 731–736). Voorburg, The Netherlands: International Statistical Institute.

Finch, S., & Cumming, G. (1998). Assessing conceptual change in learning statistics. In L. Pereira-Mendoza, L. Seu Kea, T. Wee Kee, & W. K. Wong (Eds.), *Proceedings of the fifth international conference on teaching statistics* (pp. 897–904). Voorburg, The Netherlands: International Statistical Institute.

Garfield. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, *2*, 22–38.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological issues* (pp. 311–339). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592–596.

Gliner, J. A., Leech, N. L., & 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.

Gordon, H. R. D. (2001). AVERA members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Education Research*, *26*(2.), retrieved January 12, 2006, from http://scholar.lib.vt.edu/ejournals/JVER/v26n2/gordon.html

Guzzetti, B. J., Snyder, T. E., Glass, G. V., & 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., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*(1), 1–20.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* London: Lawrence Erlbaum Associates, Publishers.

Healey, J. F. (2005). *Statistics. A tool for social research* (7th ed.). Belmont, CA: Thomson Wadsworth.

IES. (2006). *ERIC*. Retrieved November 28, 2006, from http://www.eric.ed.gov/ERICWebPortal/Home.portal?_nfpb=true&_pageLabel=Home_page [on-line].

Innabi, H. (1999). Students' judgment of the validity of societal statistical generalization. In A. Rogerson (Ed.), *Proceedings of the international conference on mathematics education into the 21st Century: Societal challenges, issues and approaches*

ISI. (2004). *Web of science*. Retrieved November 29, 2006, from http://portal.isiknowledge.com/portal.cgi?DestApp=WOS&Func=Frame [on-line].

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

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

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, *6*(1), 59–98.

Konold, C. (1991). Understanding students' beliefs about probability. In E. Von Glaserfeld (Ed.), *Radical constructivism in mathematics education* (pp. 139–156). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *44*, 1372–1381.

Krauss, S., & Wassner, C. (2002). How significance tests should be presented to avoid the typical misinterpretations. In *Proceedings of the sixth international conference on teaching statistics*. Voorburg, The Netherlands: International Statistical Institute.

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

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In *Proceedings of the sixth international conference on teaching statistics*. Voorburg, The Netherlands: International Statistical Institute.

McLean, A., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools*, *5*(2), 15–22.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher*, *29*(4), 1420.

Moore, D. S., & McCabe, G. P. (2006). *Introduction to the practice of statistics* (5th ed.). New York: W.H. Freeman and Company.

Nilsson, N. J. (1986). Probabilistic logic. *Artificial Intelligence*, *28*(1), 71–87.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 259–284). London: Lawrence Erlbaum Associates, Publishers.

Sedlmeier, P. (1998). The distribution matters: Two types of sample-size tasks. *Journal of Behavioral Decision Making*, *11*, 281–301.

Sedlmeier, P., & Gigerenzer, G. (2000). Was Bernoulli wrong? On intuitions about sample size. *Journal of Behavioral Decision Making*, *13*, 133–139.

Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 465–494). New York: Macmillan.

Shaughnessy, J. M. (2003). Research on students' understandings of probability. In J. Kilpatric, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 216–226). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M., & Dick, T. (1991). Monty's dilemma: Should you stick or switch? *Mathematics Teacher*, *84*, 252–256.

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

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Vallecillos, A. (1995). Comprensión de la lógica del contraste de hipótesis en estudiantes universitarios [Understanding of the logic of hypothesis testing amongst university students]. *Recherches en Didactique des Mathematiques*, *15*, 53–81.

Vallecillos, A. (1996). Students' conceptions of the logic of hypothesis testing. *Hiroshima Journal of Mathematics Education*, *4*, 43–61.

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. In *Proceedings of the 52 session of the International Statistical Institute* (pp. 201–204). Helsinki: International Statistical Institute. Tome 58, Book 2

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

Vallecillos, A. (2002). Empirical evidence about understanding of the level of significance concept in hypotheses testing by university students. *Themes in Education*, *3*(2), 183–198.

Vallecillos, A., & Batanero, C. (1996). Conditional probability and the level of significance in tests of hypotheses. In L. Puig & A. Gutiérrez (Eds.), *Proceedings of the 20th conference of the International Group for the Psychology of mathematics education* (pp. 271–378). Valencia, Spain: University of Valenciam.

Vallecillos, A., & Batanero, C. (1997). Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios [Activated concepts in the statistical hypotheses contrast and their understanding by university students]. *Recherches en Didactique des Mathematiques*, *17*, 29–48.

Ware, M. E., & Chastain, J. D. (1991). Developing selection skills in introductory statistics. *Teaching of Psychology*, *18*(4), 219–222.

Watson, J. M., & Moritz, J. B. (2000). Development of understanding of sampling for statistical literacy. *Journal of Mathematical Behavior*, *19*, 109–136.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, *47*, 289–312.

Wilensky, U. (1997). What is normal anyway? Therapy for epistemological anxiety. *Educational Studies in Mathematics*, *33*(2), 171–202.

Williams, A. M. (1998). Students' understanding of the significance level concept. In L. Pereira-Mendoza, L. S. Kea, T. W. Kee, & W. Wong (Eds.), *Proceedings of the fifth international conference on teaching statistics* (pp. 743–749). Voorburg, The Netherlands: International Statistical Institute.