SUITABILITY OF TEACHING BAYESIAN INFERENCE IN DATA ANALYSIS COURSES DIRECTED TO PSYCHOLOGISTS[1]

Carmen Díaz Batanero

1. INTRODUCTION

In this Thesis we focus on the use of Bayesian inference in the field of psychology from different perspectives:

1. The reflection on the current statistical practices in psychology, the reported errors and possible contribution of Bayesian inference to solve these problems. This analysis is carried out from the philosophical and psychological points of views (Chapter 1).

2. The study of some applications of Bayesian methods in psychometrics to estimate different indicators used in the Classical Tests Theory. These possibilities are analysed from the theoretical (Chapter 3) and practical (Chapters 4, 5) points of view and are applied in the process of building a questionnaire to assess conditional probability reasoning (CPR), which is also justified in the thesis.

3. The feasibility of teaching basic Bayesian elements in undergraduate psychology courses. We develop a teaching material that takes into account the previous analyses, as well as previous research in statistics education and the type of students. This material was tested with a sample of 78 students, and data on the students’ learning at the end of the experience are provided (Chapter 6).

Below we describe the aims and structure of the thesis and summarize the different studies included in the same.

[1] Abstract of the Ph. Dissertation in the Programme: Research Methods in Behavioural Sciences, Faculty of Psychology. University of Granada, Spain. Supervisor: Dr. Inmaculada de la Fuente.

2. RESEARCH AIMS AND STRUCTURE

There are four main goals in this research. For each of them we carry out one or more studies, which are related one to another as shown in Figure 1.

- *Objective 1. Rethinking the Classical Tests Theory CTT from the Bayesian point of view and analyzing the implications of this change of perspective on the estimation of some psychometric features in the tests and items.*

In Chapter 3 we analyse the implications of a Bayesian perspective on the estimation of tests mean scores, differences in mean scores, difficulty and discrimination indexes. For each of these parameters we consider both informative and non informative priors and prepare some Excel programs to carry out the computations of posterior distributions and credibility intervals. Results are useful to build and adapt other questionnaires, in particular when prior information for the psychometric features is available.

- *Objective 2. Applying the above analysis in the process of building a questionnaire and comparing results from classical and Bayesian estimates in some of the test features*.

The building of the RPC questionnaire starts from the semantic definition of the variable through content analysis of 18 statistics textbooks directed to psychology students (Studies 1-4; Chapter 4). The process follows the recommendations by APA, AERA and NCME (1999) and includes items trials, expert judgment to fix the content and select the items, pilot trial of the questionnaire and a second expert judgment to improve the items wording. Reliability and validity studies (Studies 5 and 6; Chapter 5) are carried out in different sample of students. All this process is complemented with application of Bayesian methods. The RPC questionnaire is useful to assess students' understanding of conditional probability in statistics courses and future research.

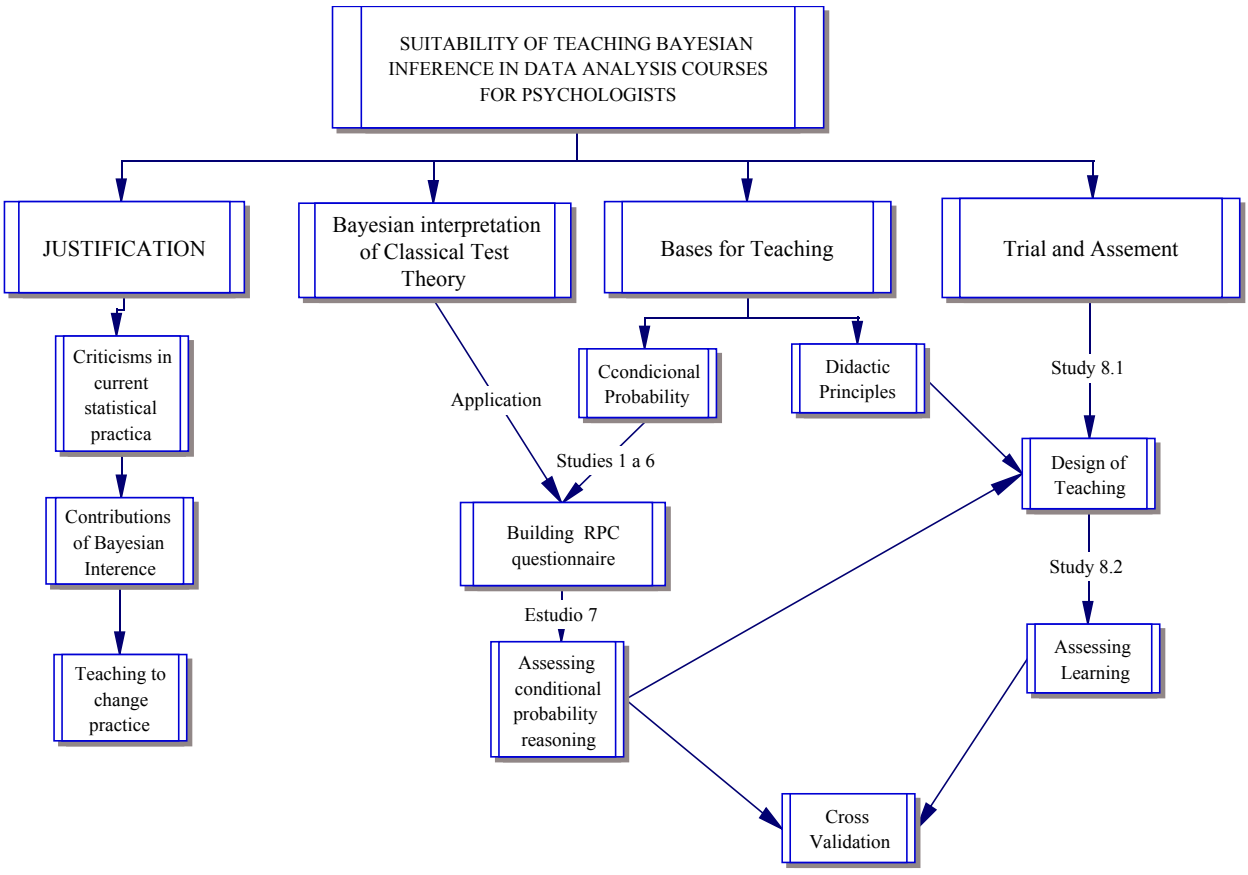
- *Objective 3. Assessing conditional probability reasoning in psychology students to decide the suitability of teaching Bayesian methods to these students.*

The RPC questionnaire is applied to a sample of 413 psychology students (Study 7) and their responses are analysed from different points of views. Students showed enough understanding of conditional probability to start the learning of Bayesian inference, but, at the same time, we found some widespread misconceptions that were taken into account in the next stage (designing a curricular proposal).

- *Objective 4. Preparing and assessing didactic materials to introduce elementary Bayesian inference to Psychology students that takes into account the previous assessment.*

The teaching materials are based on several textbooks of Bayesian inference and include activities, assessment questionnaires and Excel programs. It is available from the web page http://www.ugr.es/~mcdiaz/bayes/. An experiment is organized with a sample of 78 students (working in small groups) to try these materials (Studies 8 and 9). The posterior learning, structure of responses to assessment items and relationship with understanding conditional probability are analysed.

Figure 1. Research Structure



3. JUSTIFICATION

In Chapter 1 we present the foundations of the Thesis that can be classified in three main parts: a) Current situation in the practice of statistics inference and the need for a change; b) Possible contributions of Bayesian inference to improve the situation and need to include these methods in undergraduate courses; c) Relevance of assuring correct reasoning on conditional probability in the students before trying to teach them Bayesian inference and

need for a comprehensive questionnaire to assess this reasoning (RPC questionnaire). In the following we summarize the main points in this justification.

3.1. CRITICISMS IN THE CURRENT PRACTICE OF STATISTICS IN EMPIRICAL RESEARCH

Empirical sciences heavily rely on establishing the existence of effects using the statistical analysis of data. Statistical inference dates back almost 300 years. However, since the logic of statistical inference is difficult to grasp, its use and interpretation are not always adequate and have been criticized for nearly 50 years (for example, in Yates, 1951; Morrison & Henkel, 1970; Harlow, Mulaik & Steiger, 1997). This controversy has increased in the past ten years within professional organizations (Menon, 1993; Ellerton, 1996; Levin, 1998; Levin & Robinson, 1999; Robinson & Levin, 1997; Ares, 1999; Glaser, 1999; Wilkinson, 1999; Batanero, 2001; Fidler, 2002), which are suggesting important shifts in their editorial policies regarding the use of statistical significance testing.

Despite the arguments that statistical tests are not adequate to justify scientific knowledge, researchers persist in relying on statistical significance (Hager, 2000; Borges, San Luis, Sánchez & Cañadas, 2001; Finch, Cumming & Thomason, 2001). Some explanations for this persistence include inertia, conceptual confusion, lack of better alternative tools, and psychological mechanisms such as invalid generalization from deductive logic to inference under uncertainty (Falk & Greenbaum, 1995). Below we summarize some of the problems that were analyzed in Batanero (2000) and Díaz and De la Fuente (2004).

Common Errors in Interpreting Statistical Tests

Misconceptions related to statistical tests mainly refer to the level of significance α, which is defined as the probability of rejecting a null hypothesis, given that it is true. The most common misinterpretation of this concept consists of switching the two terms in the conditional probability. For example, Birnbaum (1982) reported that his students found the following definition reasonable: *"A level of significance of 5% means that, on average, 5 out of every 100 times we reject the null hypothesis, we will be wrong"*. Falk (1986) found that most of her students believed that α was the probability of being wrong when rejecting the null hypothesis at a significance level α. Similar results were described in Pollard and Richardson (1987), Lecoutre, Lecoutre and Poitevineau (2001) and Haller and Krauss (2002) in their studies using researchers.

Another common error is the belief in the conservation of the significance level value when successive tests are carried out on the same data set, which produces the problem of multiple comparisons (Moses, 1992). Some people believe that the p-value is the probability that the result is due to chance. The p-value however is the probability of obtaining the particular result or one more extreme when the null hypothesis is true and there are no other possible factors influencing the result. What is rejected in a statistical test is the null hypothesis, and therefore we cannot infer the existence of a particular cause in an experiment from a significant result.

Another erroneous belief is that the .05 and .01 levels of significance are justified by mathematical theory. In his book "Design of Experiments", Fisher (1935) suggested selecting a significance level of 5% as a convention to recognize significant results in experiments. In later writings, however, Fisher considered that *"in fact, no scientific worker has a fixed level of significance at which from year to year and in all circumstances, he rejects hypotheses"* (Fisher, 1956, p. 42). Instead, Fisher suggested publishing the exact p-value obtained in each particular experiment which, in fact, implies establishing the significance level after the experiment. In spite of these recommendations, research literature shows that the common arbitrary levels of .05, .01 and .001 are almost universally selected for all types of research problems and are sometimes used as criteria for publication.

Misinterpretations of the significance level are linked to misinterpreting significant results; we should distinguish between statistical and practical significance, since we might have obtained a higher level of significance with a smaller experimental effect and a larger sample size. Practical significance involves statistical significance plus a sufficiently large experimental effect.

Philosophical and Psychological Issues

Several reasons explain the difficulties in understanding statistical tests. On one hand, statistical tests involve a series of concepts such as null and alternative hypotheses, Type I and Type II errors, probability of errors, significant and non significant results, population and sample, parameter and statistics, sampling distribution. Some of these concepts are misunderstood or confused by students and experimental researchers.

Moreover, the formal structure of statistical tests is superficially similar to that of proof by contradiction. However, there are fundamental differences between these two types of reasoning that are not always well understood. In proof by contradiction we reason in the following way: If A implies B cannot happen, then, if B happens, we deduce A is false. In

statistical testing, it is tempting to apply similar reasoning as follows: If *A* implies *B* is very unlikely to happen. However, this does not imply that if *B* happens, *A* is very unlikely and here lays the confusion.

The controversy surrounding statistical inference involves the philosophy of inference and the logical relations between theories and facts. We expect from statistical testing more than it can provide us, and underlying this expectation is the philosophical problem of finding scientific criteria to justify inductive reasoning, as stated by Hume. The contribution made by statistical inference in this direction is important but it does not give a complete solution to this problem (Hacking, 1975; Seidenfeld, 1979; Cabria, 1994).

On the other hand, there are two different views about statistical tests that sometimes are confused or mixed. Fisher saw the aims of significance testing as confronting a null hypothesis with observations and for him a p-value indicated the strength of the evidence against the hypothesis (Fisher, 1958). However, Fisher did not believe that statistical tests provided inductive inferences from samples to population, but rather, a deductive inference from the population of possible samples to the particular sample obtained in each case.

For Neyman (1950), the problem of testing a statistical hypothesis occurs when circumstances force us to make a choice between two courses of action. To accept a hypothesis means only to decide to take one action rather than another. This does not mean that one necessarily believes that the hypothesis is true. For Neyman and Pearson, a statistical test is a rule of inductive behaviour, a criterion for decision-making, which allows us to accept or reject a hypothesis by assuming some risks.

The dispute between these authors has been hidden in applications of statistical inference in psychology and other experimental sciences, where it has been assumed that there is only one statistical solution to inference (Gingerenzer et al, 1989). Today, many researchers apply the statistical tools, methods, and concepts of the Neyman-Pearson theory with a different aim, namely, to measure the evidence in favour of a given hypothesis. Therefore, the current practice of statistical tests contains elements from Neyman-Pearson (it is a decision procedure) and from Fisher (it is an inferential procedure, whereby data are used to provide evidence in favour of the hypothesis), which apply at different stages of the process. We should also add that some researchers often give a Bayesian interpretation to the result of (classical) hypothesis tests, in spite of the fact that the view from Bayesian statistics is very different from the theories of either Fisher or Neyman and Pearson.

Moreover, biases in inferential reasoning can be seen simply as examples of adults' poor reasoning in probabilistic problems (Nisbett & Ross, 1980; Kahneman, Slovic & Tversky,

1982). In the specific case of misinterpreting statistical inference results, Falk and Greenbaum (1995) describe the *illusion of probabilistic proof by contradiction*, which consists on the erroneous belief that one has rendered the null hypothesis improbable by obtaining a significant result. Misconceptions around the significance level are also related to difficulties in discriminating between the two directions of conditional probabilities, otherwise known as *the fallacy of the transposed conditional* (Diaconis & Friedman, 1981). Although α is a well defined conditional probability, the expression "Type I error" is not conditionally phrased, and does not spell out to which combination of the two events it refers. This leads us to interpret the significance level as the conjunction of the two events "the null hypothesis is true" and "the null hypothesis is rejected" (Menon, 1993).

The Statistical Tests Controversy

For many years, criticisms have been raised against statistical testing, and many suggestions have been made to eliminate this procedure from academic research. However, significant results continue to be published in research journals, and errors around statistical tests continue to be spread throughout statistics courses and books, as well as in published research. An additional problem is that other statistical procedures suggested to replace or complement statistical tests (such as confidence intervals, measuring the magnitude of experimental effects, power analysis, and Bayesian inference) do not solve the philosophical and psychological problems we have described (see Fidler, 2002; Cumming, Williams & Fidler, 2004). Below we revisit some frequent criticisms that either are not justified or refer to researchers' use of statistical tests more than to the procedure itself.

Criticism 1. *The null hypothesis is never true and therefore statistical tests are invalid, as they are based on a false premise (that the null hypothesis is true).* This criticism is not pertinent because what is asserted in a test is that a significant result is improbable, given that the null hypothesis is true. This is a mathematical property of the sampling distribution that has nothing to do with the truth or falsity of the null hypothesis.

Criticism 2. *Statistical significance is not informative about the practical significance of the data, since the alternative hypothesis says nothing about the exact magnitude of the effect.* In significance testing (Fisher's approach) the aim of experimental research is directed towards theory confirmation in providing support for a substantive hypothesis and the magnitude of effect is not so important. In the context of taking a decision (Neyman- Pearson), however, the

magnitude of the effect could be relevant to the decision. In these cases, the criticism applies and statistical tests should be complemented with power analysis and/ or estimates of the magnitude of the effects (Levin, 1998; Frías, Pascual & García, 2000; Vacha-Haase, 2001).

Criticism 3. *The choice of the level of significance is arbitrary; therefore some data could be significant at a given level and not significant at another different level.* It is true that the researcher chooses the level of significance. This arbitrariness does not, however, mean that the procedure is invalid. Moreover, it is also possible, following the approach of Fisher, to use the exact p-value to reject the null hypothesis at different levels, though in the current practice of statistical testing it is advisable to choose the significance level before taking the data to give more objectivity to the decision.

Criticism 4. *Statistical significance does not provide the probability of the hypothesis being true. Nor is statistical significance informative of the true value of the parameter.* The posterior probability of the null hypothesis, given a significant result, depends on the prior probability of the null hypothesis, as well as on the probabilities of having a significant result given the null and the alternative hypotheses. These probabilities cannot be determined in classical inference. It is only within Bayesian inference that posterior probability of the hypotheses can be computed, although these are subjective probabilities (Cabria, 1994; Lecoutre, 1999; 2006).

Criticism 4. *Type I error and Type II errors are inversely related. Researchers seem to ignore Type II errors while paying undue attention to Type I error.* Though the probabilities of the two types of errors are inversely related, there is a fundamental difference between them. While the probability of Type I error α is a constant that can be chosen before the experiment is done, the probability of Type II error is a function of the true value of the parameter, which is unknown. To solve this problem, power analysis assumes different possible values for the parameter and computes the probability of Type II error for these different values.

3.2. POSIBLE CONTRIBUTIONS OF BAYESIAN INFERENCE TO IMPROVE METHODOLOGICAL PRACTICE

In this section we begin summarizing the characteristics of Bayesian inference. We then present some arguments in favour of the Bayesian methodology: a) Bayesian inference does not contain greater subjectivity than other statistical methods; b) it provides the information that researchers need and c) there is statistical software available that facilitates the

application of this methodology. We then suggest that the basic Bayesian concepts are understandable by psychology students, if a necessary didactic effort is made.

Bayesian inference

Bayesian inference is based on the systematic application of the Bayes Theorem, whose publication in 1763 disturbed the contemporary mathematicians. While in the previous conceptions of probability[2] it was assumed an objective value of probability, the possibility of revising the prior probabilities based on the new information opened by this theorem, lead to a new subjective view (Hacking, 1975; Cabriá, 1994). This new point of view also enlarges the applications of probability, since the repetition of an experience in exactly the same conditions was no more a requirement. Gradually, a distinction between frequentist probability, empirically accessible through frequencies, and epistemic probability or degree of belief in the occurrence of an event in a unique experiment (Rouanet, 1998) and two schools of inference were developed.

In Bayesian inference a parameter θ is a random variable and we associate to it a prior epistemic distribution of probabilities $p(\theta)$, which represents the knowledge (or lack of knowledge) about θ before collecting the data. Let $y = (y_1,..., y_n)$ be a data set, whose likelihood function $p(y/\theta)$ depends on the parameter, then the conditional distribution of θ given the observed data y is given by the Bayes theorem:

$$p(\theta / y) = \frac{p(y / \theta) p(\theta)}{p(y)} \quad (1)$$

In (1) $p(y) = \sum p(y/\theta)p(\theta)$, where the sum extends through the admissible range of θ (Box and Tiao, 1992; Lee, 2004). The posterior distribution $p(\theta/y)$ contains all the information about θ once the data are observed. The Bayes theorem can be successively applied in new experiments, taking as prior probabilities of the second experiment the posterior probabilities obtained in a first experiment and so on. We speak of "learning process" (Box and Tiao, 1992).

The main method in Bayesian inference is the systematic application of the Bayes theorem, and the basic aim is updating the parameters prior distributions. The posterior distribution is the essence of Bayesian estimation. The answer to the question: once we see the data, what do we know about the parameter? It is the posterior distribution, since this

[2] Classical (quotient between favourable and possible cases) and frequentist (limit of relative frequency) conceptions.

distribution synthesizes all the information about the parameter, once the data have been gathered and contains all the inferences that can be done from it (O'Hagan & Forster, 2004). The point estimate for the parameter is the mean of the posterior distribution, since it minimizes the expected quadratic error (O'Hagan & Forster, 2004). The posterior distribution will also allow us to compute the probability that the parameter is included in a given interval (credible interval) and the probability that the hypothesis is either true or false. Bayesian inference' aim is to compute the hypothesis' posterior probability, contrary to classical inference, where the hypothesis is either accepted or rejected, which is not an inference, but a decision (O'Hagan & Forster, 2004).

The predictive or marginal distribution

$$p(y) = \propto \int p(y/\theta)\, p(\theta) d\theta$$

is used to predict future values of y. It takes into account the uncertainty about the parameter value θ, as well as the residual uncertainty about y when θ is known (Lee, 2004). This kind of probability cannot be computed in classical inference (Bolstad, 2004).

Subjectivity in Bayesian methods

A fundamental difference between Bayesian and classical inference is the subjective character (not frequentist) of probabilities, since neither the problem of repeated sampling is considered nor the sample distribution is required. Subjective probabilities can be defined for any situation, whereas frequentist probabilities are only defined for events in a space sample (O'Hagan & Forster, 2004). Moreover, Bayesian methods use all the previous information available, whereas in classical inference previous information is not considered.

Since the researcher specifies the prior distribution, the Bayesian approach takes into account the researcher's perspective, his/her knowledge of the problem. There is not just one way to choose the prior distribution, which conditions the results of inference. This fact has originated strong criticisms towards Bayesian methods since they can lead researchers to obtain different results from the same data set, depending on their previous knowledge or experience. The use of non informative priors at the beginning of the application of these methods, and updating these prior distributions in new applications, with the results of the previous steps has been suggested in order to confront these criticisms.

There is also the possibility of changing the models and interpretations throughout the analysis, whereas in classical inference both hypotheses and models are settled down before gathering the data and cannot be changed. This is not reasonable, since "allowing data to

speak by themselves" is a basic idea in the mathematical modelling, where models are assumed to be useful to describe data but not to be exactly equal to data and it is therefore possible to change the model throughout the analysis (Pruzek, 1997; McLean, 2001).

The influence of prior distributions also depends on the sample size and the possible initial biases are corrected in successive experiments, since the weight lays on the likelihood as the sample size is progressively increased (Lindley, 1993). It is also advisable to repeat the analysis with different priors and inform about the differences obtained in the posterior distributions (Zhu and Lu, 2004). Procedures are standardised, using conjugated distributions, so that both the prior and posterior distribution belongs to the same functions family (Cabriá, 1994).

On the other hand, frequentist methods are not free of subjectivity: the significance level is arbitrarily defined, so that the same data is statistically significant or not depending on the chosen significance level (Skipper, Guenter & Nash, 1970). Statistical significance has no sense when the sample size is so big that any detected difference led to rejecting the null hypothesis. The variable definition, scale of measurement, significance tests used, are other subjective choices and even more, subjectivity is unavoidable in the interpretation of the results (Ayçaguera & Benavides, 2003). Of course, subjectivity does not imply arbitrairity; it is inevitable in social sciences due to the inherent randomness in its variables and has an important paper in the scientific research. The scientific community accepts the different findings, by establishing methodological or plausibility criteria (Matthews, 1998).

What are the Bayesian answers to researcher's needs?

Several works suggest that Bayesian inference provides a better answer to the researcher's needs as compared with frequentists inference (Lindley, 1993; Lecoutre, 1999; 2006).

Firstly, the meaning of probability in Bayesian statistics is identical to that of ordinary language: *conditional measurement of uncertainty* associated to the occurrence of an event, when some assumptions are assumed (Bernardo, 2003). This is the intuitive - although incorrect - interpretation that many scientists give to the frequentist probabilities associated to hypotheses tests, whose results are unconsciously interpreted in Bayesians terms (Falk, 1986; Gingerenzer, 1993; Rouanet, 1998; Lecoutre, 2006; Lecoutre, Lecoutre & Poitevineau, 2001; Haller & Krauss, 2002).

Consequently, the Bayesian interpretation of inference is simpler and more natural than that of frequentists inference (Pruzek, 1997), besides providing a base for coherent decision making in uncertainty situations (Western, 1999). In addition, Bayesian inference provides a

totally general method, because its application does not require a particular kind of distribution and sampling distributions do not need to be deduced (Bernardo, 2003). Next we analyze the Bayesian answer to several questions of interest for researchers.

Effect size

A recommendation to complement hypothesis tests is to study the effect size, but a point estimation is insufficient, since it does not consider the sample error (Poitevineau, 1998). A power study would be recommendable to avoid erroneous conclusions about the absence of an effect when the result is nonsignificant (Cohen, 1990), but power computations does not depend on the statistical value observed in the sample and is therefore not pertinent to interpret a particular result, once the data are gathered (Falk & Greenbaum, 1995). Confidence intervals have the same frequentist interpretation than hypotheses tests, since they only indicate the proportion of intervals with a given sample size computed from the same population that would cover the parameter value, but they do not give information about whether the calculated interval covers the parameter or not (Cumming, Williams & Fidler, 2004).

Effect sizes and their magnitude appear in natural way in the Bayesians methods, which consider the parameter as a random variable. The probability that this parameter takes a certain value can be computed via the posterior distribution; for example it is possible to use sentences such as "the probability that the effect is larger than *a* is equal to 0.25". The credibility interval also provides the limits in which the parameter is included with a certain probability (Poitevineau, 1998; Lecoutre, 2006).

Hypothesis tests

The p-value provides a probability that is not useful for researchers: the probability of collecting data more extreme than the obtained if we repeated many times the experiment and the hypothesis were true (Matthews, 1998). But no researcher is interested in repeating the same experiment indefinitely and the aim of the scientific research is not to make a decision about the certainty of the hypothesis but adjusting our degree of belief in the hypothesis that is being tested (Rozeboom, 1970).

Interpreting the rejection of the null hypothesis as direct support to the research hypothesis (alternative) is incorrect, since a significant result does not indicate the magnitude of the effect, so that the statistical hypothesis does not inform on the practical meaning of the data (Hager, 2000; Finch, Cumming & Thomason, 2001). This can produce situations in

which rejecting a null hypothesis does not provide any new information, since the only thing we can deduce when we reject a hypothesis is that there is an effect, but not its direction or magnitude (Falk & Greenbaum, 1995; Lecoutre, 1999).

On the contrary, in Bayesian inference we can compute the hypothesis posterior probabilities and the probabilities that the effect has a given size (Lindley, 1993). Moreover, the Bayesian method is comparative. It compares the probability of the observed event under the null hypothesis and under different alternative hypotheses (Lindley, 1993). Besides, in some situations, as bioequivalence tests, the interest is centred in verifying the null hypothesis, that is, we hope the treatments are equivalent (Molinero, 2002). In these cases the Bayesian approach is much more natural than the frequentists one, since we try to accept (not to reject) the null hypothesis.

Predictive probabilities and replication

Interpreting statistical significance as support to data replicability does not have a statistical base (Falk, 1986; Gingerenzer, 1994; Cohen, 1994; Falk & Greenbaum, 1995; Pascual, García & Frías, 2000). Statistical significance neither can be taken as an evidence that the research hypothesis is true; nor it provides the probability of the hypothesis; there is therefore no base to study replication and it does not provide verifiable evidence to replication either (Sohn, 1998).

In the Bayesian approach we can compute the probability of a future event, using the predictive distribution, which is given by the denominator in the Bayes formula, that is, the weighted average of the probability function, weighted by the prior probabilities (Berry, 1995). This distribution serves to study the possibility of replication of our results or to compute the sample size needed for a future study to be conclusive (Lecoutre, 1996). Of course, in case the requirements of data precision and sound procedures are fulfilled (Sohn, 1998). Correctly understood, replicability is related to the data reliability and consistency, and the only way to achieve it is sucessive empirical trials (Pascual, García & Frías, 2000).

Use of previous information

Whereas frequentist methods consider each sample as completely new and do not incorporate the information of previous studies, in the Bayesian framework we conceive a sequence of articulated experiments, where the information of each of them is used in the following step (Pruzek, 1997); the possibility of different opinions or knowledge is also accepted (Lindley, 1993). Although is possible to use Bayesian inference when there is no previous

information about the parameter, the most interesting characteristic is the use "informative" priors whenever this is possible, or even investigate the effect of different priors. The central idea of Bayesian approach is updating the probabilistic knowledge about the phenomenon, based on the information available.

Computational viability of Bayesian methodology

A requirement to introduce new data analysis methods is the availability of calculation programs that facilitate their application. In the last years several researchers are developing diverse Bayesian programs, so that this approach is being introduced gradually in Social Sciences. For example, Albert (1996) published some Minitab subroutines for elementary Bayesian analysis that can be downloaded from the author's website (http://bayes.bgsu.edu/).

First Bayes (http://www.tonyohagan.co.uk/1b/) was prepared at Sheffield University to teach elementary Bayesian concepts. It admits different families of distributions and calculates posterior and predictive probabilities in uniparametrics models, analysis of variance and regression (Lawrence, 2003).

PAC (Lecoutre, 1996) also allows the analysis of data from general experimental designs, incorporating univariate and multivariate variance analysis, including repeated and covariable measures. The program includes frequentist and Bayesian analysis, with prior informative and non informative. It was developed by a research group that tries to incorporate Bayesian analysis in the statistical methods more frequently used in psychology. A reduced version is freely distributed from the group website (http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.html).

For more complex analyses Buggs (Bayesian inference Using Gibbs Sampling) is an interactive and flexible software Windows compatible, that allows complex Bayesians calculations, based on simulation (see in http://www.mrc-bsu.cam.ac.uk/bugs/). There are on line facilities, such as tutorial, user groups and examples.

BACC (Bayesian Analysis Computation and Communication) was developed from a project funded by the National Science Foundation in the United States, and offers resources for Bayesian calculations, freely available. The emphasis is put in the combination of models and the development of predictive distributions. There are versions available for Matlab, S-PLUS and R, for Windows, UNIX and Linux systems (http://www2.cirano.qc.ca/~bacc/).

Other Bayesian computation programs, some specific are listed in http://www.mas.ncl.ac.uk/~ndjw1/bookmarks/Stats/Software-Statistical_computing/Bayesian_software/index.html/.

Didactic viability of elementary Bayesians methods

Introducing a new methodology in psychology will require its understanding by the possible users, that is to say, will depend on the degree to which we are able to transmit its main ideas in applied statistic courses. Iglesias et al. (2000) suggest the following content to introduce Bayesian inference, along with classical inference in undergraduates' courses following the approach by De Groot (1988):

- Basic concepts: population, parameter, sample, statistics, likelihood function, prior and posterior distributions.
- Point estimation: Classical and Bayesian methods.
- Interval estimation: Confidence and credibility intervals.
- Hypothesis tests: Classical and Bayesian tests, multiple decision problems.

In this sense, we found a increasing number of textbooks whose understanding does not require much mathematical knowledge and where basic Bayesian inference elements are contextualized in examples interesting and familiar for the students (for example Berry, 1995 or Albert & Rossman, 2001). These materials can be complemented with many references that explain in a simple way the basics of Bayesian inference (e.g. Ayçaguera & Benavides, 2003; Ayçaguera & Suárez, 1995). We can also find Internet didactic resources that facilitate the learning of these concepts, such as applets that visualize the Bayes theorem or the probability distributions, or compute posterior distributions, inference for means and proportions with discrete or continuous prior distributions (see, for example Jim Albert site, http://bayes.bgsu.edu/).

Most of the authors mentioned in this section have incorporated Bayesian methods to their teaching and have reported that students seem to understand better Bayesian inference than classical inference. We also found descriptions of concrete teaching experiments and suggestions about the way to carry them out (Bolstad, 2002). We are conscious, nevertheless, that this position is still controversial (e.g. Moore, 1997) due to the scarce empirical research on the students learning within statistics courses. Moreover, biases in conditional probability reasoning, as described below, may affect students' learning of Bayesian inference.

3.3. CONDITIONAL REASONING AND ITS RELEVANCE FOR UNDERSTANDING BAYESIAN INFERENCE

Research on understanding conditional probability has been carried out with both secondary school and University students. Fischbein and Gazit (1984) organized teaching experiments with 10-12 year-olds and found that conditional probability problems were harder in without replacement situation as compared to with replacement problems. Following that research Tarr and Jones (1997) identified the following four levels of thinking about conditional probability and independence in middle school students (9-13 year-olds):

- Level 1 (subjective): students ignore given numerical information in making predictions.
- Level 2 (transitional): students demonstrate some recognition of whether consecutive events are related or not; however, their use of numbers to determine conditional probability is inappropriate.
- Level 3 (informal quantitative): students' differentiation of "with and without replacement situations" is imprecise as is the quantification of the corresponding probabilities; they are also unable to produce the complete composition of the sample space in judging independence.
- Level 4 (numerical): students state the necessary conditions for two events to be related, they assign the correct numerical probabilities and they distinguish between dependent and independent events in "with (e.g. item 15 in appendix) and without (items 4, 9) replacement situations".

Even when students progress towards the upper level in this classification (see also Tarr & Lannin, 2005), difficulties still remain at high school and University. This is shown in the various studies we summarize below, from which we have taken some of the items in our questionnaire. The full questionnaire is included in Appendix 1.

Conditioning and causation

It is well known that if an event B is the cause of another event A whenever B is present A is also present and therefore $P(A/B)=1$. On the contrary $P(A/B)=1$ does not imply that B is a cause for A, though the existence of a conditional relationship indicates a possible causal relationship. From a psychological point of view, the person who assesses the conditional probability $P(A/B)$ may perceive different type of relationships between A and B depending on the context (Tversky & Kahneman, 1982a). If B is perceived as a cause of A, $P(A/B)$ is

viewed as a causal relation, if *A* is perceived as a possible cause of *B, P(A/B)* is viewed as a diagnostic relation. At other times people confuse the two probabilities *P(A/B)* and *P(B/A)*; this confusion was termed the *fallacy of the transposed conditional* (Falk, 1986). Item 10 in Appendix 1 was included to assess these difficulties.

Causal reasoning and the fallacy of the time axis

Falk (1989) gave item 17 in the Appendix 1 to 88 university students and found that while students easily answered part (a), in part (b) they typically argued that the result of the second draw could not influence the first, and claimed that the probability in Part B is 1/2. Falk suggested that these students confused conditional and causal reasoning and termed *fallacy of the time axis* their belief that an event could not condition another event that occurs before it. This is a false reasoning, because even though there is no causal relation from the second event to the first one, the information in the problem that the second ball is red has reduced the sample space for the first drawing. Hence, *P (B1 is red/ B2 is red) =1/3*. Similar results were found by Gras and Totohasina (1995) who identified two different misconceptions about conditional probability in a survey of seventy-five 17 to 18 year-old secondary school students:

- The *chronological conception* where students interpret the conditional probability *P(A/B)* as a temporal relationship; that is, the conditioning event *B* should always precede event *A*.
- The *causal conception* where students interpret the conditional probability *P(A/B)* as an implicit causal relationship; that is, the conditioning event *B* is the cause and *A* is the consequence.

Synchronical and diachronical situations

Another issue involving time and conditional probability has been identified in the literature. In *diachronical* situations (e.g. items 5 and 17 in the Appendix) the problem is formulated as a series of sequential experiments, which are carried out over time. *Synchronical* situations (e.g. items 4, 8 and 10 in the Appendix) are static and do not incorporate an underlying sequence of experiments. Formally the two situations are equivalent, however Sánchez and Hernández (2003) found that students did not always perceive the situations as equivalent and produce additive solutions to synchronical conditional problems.

Solving Bayes problems

As regards Bayesian reasoning (see a summary in Koehler, 1996), early research by Tversky and Kahneman (1982a) suggests that people do not employ this reasoning intuitively and establish the robustness and spread of the base-rate fallacy in students and professionals (Bar-Hillel, 1983). Totohasina (1992) suggested that part of the difficulty in solving Bayes' problems is due to the representation chosen by the student to solve the problems and that using a two way table is an obstacle to perceive the sequential nature of some problems, and therefore can lead students to confuse conditional and joint probability.

Recent research suggests that Bayesian computations are simpler when information is given in natural frequencies, instead of using probabilities, percentages or relative frequencies (Cosmides & Tooby, 1996; Gigerenzer, 1994; Gigerenzer & Hoffrage, 1995). The reason is that natural frequencies (absolute frequencies) correspond to the format of information humans have encountered throughout their evolutionary development. In particular, Bayes problems transform to simple probability problems if the data are given in an adequate format of absolute frequencies. Sedlmeier (1999) analyzes and summarizes recent teaching experiments carried out by psychologists that follow this approach and involve the use of computers. The results of these experiments suggest that statistical training is effective if students are taught to translate statistical tasks to an adequate format, including tree diagrams and absolute frequencies (Martignon & Wassner, 2002).

Other difficulties and need for a comprehensive assessment questionnaire

Other difficulties include problems in defining the conditioning event (Bar-Hillel & Falk, 1982) and misunderstanding of independence (Sánchez, 1996; Truran & Truran, 1997). People also have problems with compound probabilities. Kahneman and Tversky (1982a) termed conjunction fallacy people's unawareness that a compound probability cannot be higher than the probability of each single event.

The previous study of literature showed us that there is a large amount of research on this topic but we found no comprehensive questionnaires to globally assess students' understanding and misconceptions on these topics and relate one to another. As a result, one of the goals in this research was constructing a questionnaire, which takes into account the content of conditional probability taught in the Spanish universities to psychology students, as well as the biases and misconceptions described in the literature. Studies 1-6 were oriented to construct and validate the questionnaire; Study 7 was directed to assess conditional reasoning with this questionnaire in a sample of 414 psychology students after teaching of the topic. We

also analyse possible relationships between formal knowledge of the topics and psychological biases (Study 6) and relationship between understanding conditional probability and learning Bayesian inference (Study 9). Even when we focus on psychology students, the questionnaire is useful in assessing conditional probability reasoning for other undergraduate or high school students.

4. A BAYESIAN APPROACH TO CLASSICAL TESTS THEORY

In the Classical Tests Theory (Muñiz, 1994; Martinez Arias, 1995), formulated by Spearman (1904), the empirical score X obtained by a subject in a test is a random variable and it is made of two components: the subject's true score (V) in that test, that it is assumed to be constant and the error measurement (e). The model makes the following hypotheses (Muñiz, 1995):

$$X=V+e$$

$$E(X)=V$$

- $E(e_i;)=0$, for the population of subjects being measured, as well as for the infinite repetitions of the test in a subject. It is supposed that errors follow a normal distribution.
- $\rho(V,e)=0; \rho(e_i,e_j)=0$. It is assumed that the measurement error is not correlated with the true score and the measurement errors of different subjects are also independent.

In a consistent Bayesian formulation of the Classical Tests Theory, the basic assumptions should be respected and the main difference is considering the model parameters as random variables, with prior and posterior distributions. Accepting this assumption, the estimation of these parameters should be carried out with a Bayesian methodology, following its procedures and objectives. Consequently, the true score is now a random variable with a normal prior distribution[3]. From these assumptions we derive the following equalities, similar to those in CTT since they are still applicable when V is a random variable:

$$E(X)=E(V)$$

$$\sigma_X^2=\sigma_V^2+\sigma_e^2$$

$$\rho_{XV}^2=\frac{\sigma_V^2}{\sigma_X^2}=1-\frac{\sigma_e^2}{\sigma_X^2}=1-\rho_{Xe}^2$$

[3] Since the true score is sum of scores in the different items, approximated normality is reasonable.

Mean score

We can use Bayesian inference to estimate the population mean, or the difference of two different means, with both informative and non informative priors. For non informative prior distribution two cases appear:

- The standard deviation σ_i of the average prior distribution is known. In this case, for a uniform prior distribution, the average posterior distribution is normal $N(\bar{x}, \sigma/\sqrt{n})$ where $\bar{x}$ is the sample mean. The equation $Z = \frac{\mu_f - \bar{x}}{\sigma / \sqrt{n}}$ follows a distribution N (0,1) (Berry, 1995). The point estimator of the mean on the posterior distribution μ_f is the sample mean $\bar{x}$ of the data. The credibility interval for a credibility coefficient α is given by: $(\bar{x} - Z_{1-/1.\alpha)/2}\sigma/\sqrt{n}; \bar{x} + Z_{1-/1.\alpha)/2}\sigma/\sqrt{n})$, being Z a percentile of the standard normal distribution.

- If σ (population standard deviation) is not known, we can use *s*, the unbiased estimation of the standard deviation (sample cuasivariance square root) and the *T* distribution with *n-1* degrees of freedom, being *n* the sample size of data (Bolstad, 2004).

- When the prior distribution for the population mean follows a normal distribution $N(\mu_i, \sigma_i)$ and the standard deviation σ_i on the prior distribution of the mean is known, the posterior distribution also follows a normal distribution $N(\mu_f, \sigma_f)$. The values of the mean and standard deviation of the posterior distribution are given by the following formulas:

$$\mu_f = \frac{\frac{n\bar{x}}{s^2} + \frac{\mu_i}{\sigma_i^2}}{\frac{n}{s^2} + \frac{1}{\sigma_0^2}} \qquad \sigma_f = \frac{1}{\sqrt{n/s^2 + 1/\sigma_i^2}}$$

In previous expressions *n* is the sample size, $\bar{x}$ and *s* the mean and standard deviation of the sample. For the case that the standard deviation σ_i in the prior distribution of the mean is not known, this one is estimated from the square root of sample cuasivariance *s*. The previous formulas of the mean and standard deviation of the posterior distribution are the same, but now the distribution will be *T* with *n-1* degrees of freedom (being *n* the sample size), that can be approximated to the normal distribution with a sufficient sample size (Bolstad, 2004).

Difference of two mean scores

The commonest situation is the comparison of two independent simples, where different cases can be found. We will only deal with the case of prior informative distributions, since the non informative case can be included in this one.

Case 1. Identical known variances. The mean and variance of the score difference in the posterior distribution are given by:

$$\mu_d^f = \mu_1^f - \mu_2^f$$

$$\sigma_d^{2f} = \sigma_1^{2f} + \sigma_2^{2f}$$

which would coincide with the mean and variance of the sample distribution in the case of non informative prior distribution. The credibility interval of the means difference for a α credibility coefficient would be:

$$(\mu_1^f - \mu_2^f \pm Z_{1-/1.\alpha)/2}\sqrt{\sigma_1^{2f} + \sigma_2^{2f}})$$

Case 2. Different known variances. When and prior distributions are independent in both samples, posterior distributions will be also independent. The mean and variance of the posterior distribution will be again[4]:

$$\mu_d^f = \mu_1^f - \mu_2^f$$

$$\sigma_d^{2f} = \sigma_1^{2f} + \sigma_2^{2f}$$

and the credibility is given by:

$$(\mu_1^f - \mu_2^f \pm Z_{1-/1.\alpha)/2}\sqrt{\sigma_1^{2f} + \sigma_2^{2f}})$$

Case 3. Variances are not known. In this case, each of the variances should be estimated from the sample data (using the sample cuasivariances $s_1^2; s_2^2$). This increases the uncertainty of the estimation, and therefore a *T* distribution will be used, instead of the normal distribution (Box & Tiao, 1992). The degrees of freedom are given by the Satterhwaite formula: (Bolstad, 2004):

$$\upsilon = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1}}$$

[4] In this case the initial variances are different.

The approximated credibility interval is given by

$$(\mu_1^f - \mu_2^f \pm T_{1-/1.\alpha)/2}\sqrt{\sigma_1^{2f} + \sigma_2^{2f}})$$

where mean and variance on the posterior distributions are given by (1), the prior variances are estimated by the sample cuasivariances and the Satterhwaite formula is used for calculating the degrees of freedom. For the non informative prior distribution case, this expression is:

$$(\mu_1^f - \mu_2^f \pm T_{1-/1.\alpha)/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})$$

which coincides with the frecuentist confidence interval, but with a different interpretation.

Estimation of difficulty indexes

The difficulty index is defined as the proportion *p* of subjects that will get right the item, between all those that try to solve it in a certain population (Thorndike, 1991). Whereas in classical inference, the proportion *p* is considered constant, in Bayesian inference the difficulty index *p* is a random variable. Given a prior probability function *Be(a,b)* for a proportion, if in a new sample we observe *e* successes and *f* failures, the posterior probability function is *Be(a+e, b+f)* (Serrano, 2003).

Any Beta with *a=b* can be used as non informative prior, that is to say, a uniform distribution of the parameter *p* (Lecoutre, 1996). In our study we use *Be(0.5,0.5)* as recommended by Lecoutre (1996) or Serrano (2003). The credibility interval is given by:

$$\left[\beta^{-1}_{0,5+p,\ 0,5+q}(\alpha/2) - \beta^{-1}_{0,5+p,\ 0,5+q}(1-\alpha/2)\right]$$

where $\beta^{-1}_{a,b}$ is the *Beta (a,b)* distribution inverse function and α the credibility coefficient.

Estimation of discrimination indexes

A first approach to study the discrimination indexes is analysing the difference in the proportion of the item success in two groups of students with different competence[5]. Let p_s and p_i be the difficulty indexes in the higher and lower groups. In the classical theory these parameters are unknown constants in their respective populations and the point estimation of the discrimination index is:

$$d = \hat{p}_s - \hat{p}_i$$

[5] For example, students with and without instruction.

where $\hat{p}_s$ $\hat{p}_i$ are the point estimators of p_s and p_i respectively. In the Bayesian interpretation, the previous proportions and their difference would be random variables. If the prior distribution for p_s p_i are taken from the Beta family, we will obtain a posterior Beta distributions for each of these proportions. Since the populations are independent, the posterior joint distribution of the bidimensional variable (p_s p_i) is the product of two posterior distributions for each proportion.

In case of non informative prior (for example, *B(1,1)*), let e_s be the successes and f_s the failures in the higher group and e_i the successes and f_i the failures in the lower group. The respective estimators for the proportions are:

$$\hat{p}_s = \frac{e_s + 1_s}{2 + e_s + f_s} \qquad \hat{p}_i = \frac{e_i + 1}{e_i + f + 2_i}$$ (Albert, 1995; 1996)

Let the prior distribution for p_s be *B* (a_s, b_s) and the prior distribution for p_i *B* (a_i, b_i). If we achieve e_s successes and f_s failures in the higher group and e_i successes and f_i failures in the lower group, the respective estimators of the proportions are:

$$\hat{p}_s = \frac{a_s + e_s}{a_s + b_s + e_s + f_s} \qquad \hat{p}_i = \frac{a_i + e_i}{a_i + b_i + e_i + f_i}$$ (Albert, 1996)

In both cases the posterior distributions of the populations p_s is *B* (a'_s, b'_s) and that of p_i is *B* (a'_i, b'_i), that will be given by the previous formulas and are independent (Bolstad, 2004). Following Berry (1995) the estimators for the means in the posterior distribution are:

$$\hat{p}_s = \frac{a'_s}{a'_s + b'_s} \qquad \hat{p}_i = \frac{a'_i}{a'_i + b_i'}$$

The estimators for the standard deviations in the posterior distributions will be (Bolstad, 2004):

$$\partial_s = \sqrt{\frac{\hat{p}_s \hat{q}_s}{n_s + 1}} \quad \partial_i = \sqrt{\frac{\hat{p}_i \hat{q}_i}{n_i + 1}}$$

Consequently, the difference of proportions is approximately a normal distribution $N(\hat{p}_s - \hat{p}_i, \sqrt{\partial_s^2 + \partial_i^2})$, so that the approximated credibility interval is given by:

$$\hat{p}_s - \hat{p}_i \pm Z_{1-\alpha/2}^{-1} \sqrt{\partial_s^2 + \partial_i^2}$$

where *Z* is the normal *N(0,1)* distribution.

Estimating correlations and reliability coefficients

There are diverse procedures to estimate the reliability coefficient, some of which are based on estimating the correlation coefficient between scores in two administrations of the questionnaire or between scores in two equivalent forms of the questionnaire: test-retest; parallel forms and split-half reliability. In estimating these coefficients and other psychometric features[6] the correlation coefficient is used, which is a random variable in the Bayesian interpretation. Given a set of observed pairs $(x_1, y_1), (x_1, y_1), \ldots (x_1, y_1)$ for a bidimensional random variable (X, Y) with bivariate normal distribution, let's assume that the mean, variances and correlation of the scores are given by:

$$E(X) = \mu; Var(X) = \sigma^2$$
$$E(Y) = \eta; Var(X) = \varphi^2$$
$$\rho(X, Y) = \rho$$

Assume we have computed the means $\bar{x}$ and $\bar{y}$ and correlation r in the data. In case of non informative priors for the means and variances of X and Y, and given a prior distribution for the correlation coefficient $p(\rho)$, a reasonable estimation for the correlation coefficient posterior distribution is given by (Lee, 2004):

$$p(\rho/(x, y)) \propto p(\rho)\frac{\left(1-\rho^2\right)^{(n-1)/2}}{\left(1-\rho r\right)^{n-3/2}}$$

Replacing $\rho = \tanh\xi; r = \tanh z$, a new estimation is obtained, this time through the normal distribution:

$$\xi \sim N(z, 1/n)$$

This approximation can be used to find credibility intervals for the hyperbolic tangent of the correlation coefficient and from these intervals, inverting the change of variable; we find the interval for the correlation coefficient.

For informative priors, let's assume that in the first occasion we observe a correlation coefficient r_1 in a sample size n_1, which lead to a posterior distribution N($tanh^{-1}$ r_1, $1/n_1$). In a second occasion we observe a correlation coefficient r_2 in a sample size of n_2. When taking the posterior distribution in the first observation as a prior distribution in the second experiment, we can apply the formulas for estimating the mean of the normal distribution. Therefore, to estimate $tanh^{-1}$ r we have a normal posterior distribution, whose mean and

[6] E.g. the discrimination index can also be assessed as correlation between the item score and total score in the test.

variance are given by the following expressions:

$$Variance = \frac{1}{n_1 + n_2}$$

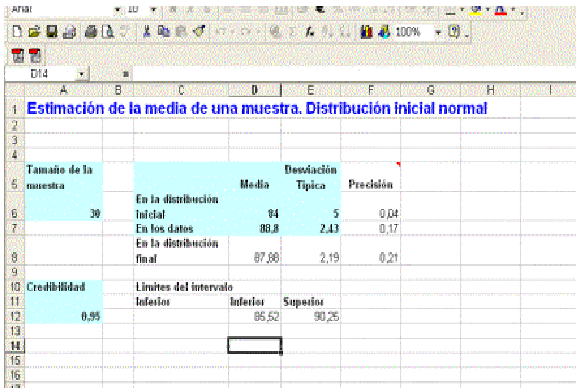
$$Mean = Variance(n_1 \tanh^{-1} r_1 + n_1 \tanh^{-1} r_1)$$

Again this transformation is applied to obtain a credibility interval of the hyperbolic tangent arc for the correlation coefficient, and inverting the transformation we obtain the credibility interval for the correlation coefficient.

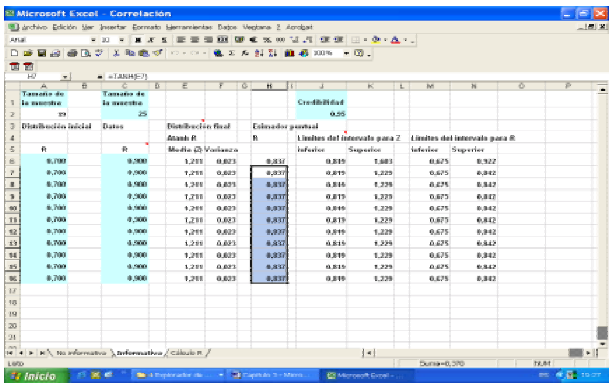
Computation software

In order to make the above calculations we prepare a set of Excel programs (See examples in Figure 2), using the formulas given in the previous sections, for each of the cases described. We also have distinguished (in different sheets of Excel files) the informative and non informative prior cases. The programs permit the variation of credibility and confidence coefficients, sample sizes, prior distributions parameters, sample statistics etc. The data statistics required can be computed with SPSS or another statistical program.

Figure 2. Some Excel programmes developed

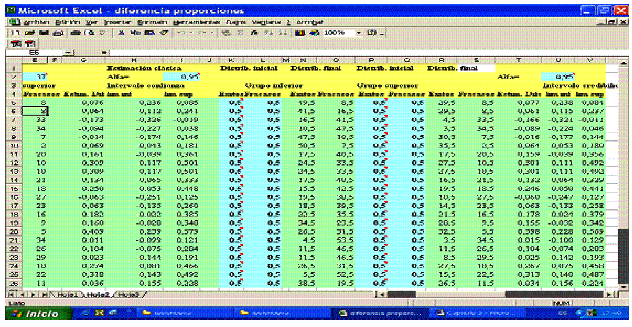


Mean



Correlation

Difference of means



Difference of proportions

In summary the above analysis was carried out to follow the Research Objective 1: *Rethinking the Classical Tests Theory (TCT) from the Bayesian point of view and analyzing the implications of this change of perspective on the estimation of some psychometric features in the tests and items.*

5. BUILDING AND VALIDATING THE CPR QUESTIONNAIRE

The objective 2 in this research *was to apply the above analysis in the process of building a questionnaire and compare results from classical and Bayesian estimates in some of the test features.* At the same time Objective 3 was *assessing conditional probability reasoning in psychology students to decide the suitability of teaching Bayesian methods to these students*. In Chapters 4 and 5 of the thesis we describe the process of building and validating the CPR questionnaire with the purpose of fulfil these two aims.

The instrument should be useful to assess in just one application the biases and misunderstanding related to conditional probability described in previous research and summarized in section 3.3 in addition to the conceptual and procedural knowledge included in the teaching of the topic in the training of psychologists in Spain. Below we briefly describe the process of building the questionnaire which is explained in detail in Chapters 4 and 5 of the thesis. This procedure includes the use of Bayesian methods to estimate difficulty and discrimination indexes (both as difference in averages and as item- total correlation, test-retest and split-half reliability coefficients) at different stages in the process. We use non informative priors in the first application of each estimation procedure; in next steps the previous final distributions are used as new informative priors.

Steps in building the questionnaire

The building of CPR questionnaire was based on a rigorous methodological process, which included the following steps:

1. *Semantic definition of the variable* (Study 1). In educational measurement (e.g. Millman & Greene, 1989) a distinction is made between constructs (unobservable psychological traits, such as *understanding of conditional probability*) and the variables (e.g. score in a questionnaire) we use to make inferences regarding the construct. In order to achieve objectivity in defining our variable, we decompose the construct "understanding conditional probability" in semantic units. These semantic units were defined after a content analysis of 19 text books used in the teaching of statistics to psychologists. The

conditional probability content in the textbooks was analysed and the definitions, properties, relationships with other concepts and procedures were classified in a reduced number of categories by means of a systematic and objective identification (Ghiglione & Matalón, 1991). To select the books, the list of references recommended in statistics courses was requested to the 31 Faculties of Psychology in Spain. All the textbooks recommended by at least 4 different Universities were analysed, after discarding some books in which conditional probability was not included.

2. *Constructing an item bank*. The aforementioned analysis was complemented with our revision of previous research on conditional probability reasoning, that also served to compile a sample of n=49 different items used in this research, some of which had been used by different authors. These items were translated into Spanish and reworded to make their format homogeneous and improve their understanding.
3. *Selection of items* (Study 2). The item difficulty (percentage of correct answers) and discrimination (correlation with test total score) were estimated from the answers by different samples of psychology students (between 49 and 117 students answered each pilot item) by classical and Bayesian procedures. Final selection of items took into account these two parameters as well as results from expert judgment. Ten statistics education researchers from five different countries (Brazil, Colombia, Mexico, Spain and Venezuela) who had themselves carried out research related to conditional probability or independence were asked to collaborate. They were asked to value (in a 5-point scale) the adequacy of the content units to understanding conditional probability as well as the suitability of each item to assess understanding for each specific content unit. The final items in the questionnaire were selected in such a way that a) the intended content of the questionnaire was covered (see Table 1); b) there was an agreement from the experts about the item adequacy; and c) item difficulty and discrimination were suitable.
4. *Formatting and revising the items.* We included two different formats: a) Multiple choice items with 3-4 possible responses were used to allow quick evaluation in the sample of some of the most pervasive biases described in the previous literature (e.g. item 3 taken from Tversky and Kahneman (1982a) which evaluates the base-rate fallacy, item 5 taken from Sánchez (1996) assesses the confusion between independent and mutually exclusive events and item 9 taken from Tversky and Kahneman (1982b) assesses the conjunction fallacy); b) Open- ended items were also used to better understand students' strategies in problem solving (e.g. item 16) and their understanding of definitions and properties (e.g., items 1, 2).

5. *The pilot trial of the instrument* (Study 3) took place in the academic year 2003-2004 with a small sample of *n=57* Psychology students in order to make a preliminary estimation of the questionnaire reliability and validity. A second sample of *n=37* students majoring in Mathematics was used to compare the performances in the two groups and to identify items with and without discriminative properties. Classical and Bayesian estimates of items difficulties and discrimination (both as item-total correlation and as difference of averages) were provided. A first estimation of internal consistency reliability provided a value Alfa= 0.787. Content validity was assessed through content analysis of items in the pilot questionnaire and through expert judgment of both content units and fitting of items to assess each content unit.
6. *Revising the pilot questionnaire* (Study 4). After discarding those items with bad psychometric features, a new expert judgment served to improve the wording of the items. Thirteen expert methodology instructors were given three alternative wordings for each item and were asked to order the three versions, as regards methodology standards, as well as give the reasons for their choice. Rank statistics were used to summarise the data. Non parametric tests (Kendall & Friedman) showed clear agreement in the option selected by the experts for each item. This version was included in the final questionnaire and additional suggestions by the methodology instructors were used to still improve readability.

Table 1. Primary content assessed by each item

Content	Item
1. Defining conditional probability; giving appropriate examples	1
2. Recognising that a conditional probability involves a restriction in the sample space	2
3. Base rates fallacy	3
4. Distinguishing conditional, simple and joint probabilities	6
5. Distinguishing a conditional probability and its inverse (transposed conditional fallacy	6
6. Conjunction fallacy	9
7. Distinguishing independent and mutually exclusive events	4
8. Computing conditional probabilities in a single experiment	8
9. Solving conditional probability problems in a sampling with replacement setting	12
10. Solving conditional probability problems in a sampling without replacement setting	5
11. Computing conditional probabilities from joint and compound probabilities	7
12. Solving conditional probability problems when the time axis is reverted	17
13. Distinguishing conditional, causal and diagnosis situations	10
14. Solving conditional probability problems in a diachronic setting	14
15. Solving conditional probability problems in a synchronic setting	15
16. Solving compound probability problems by applying the product rule to independent events	13
17. Solving compound probability problems by applying the product rule to dependent events	18
18. Solving total probability problems	11
19. Solving Bayes problems	3, 16

The final questionnaire (see Appendix 1) is composed by 18 items, with some sub-items, which score independently and some open-ended items. In Table 1 we present the items primary contents that cover the content in the books analysed as well as main biases described in the literature. There is one item covering each content (item primary content); additionally each item also assesses some other secondary contents (described in detailed in Study 3).

CPR reliability

Once the questionnaire was finished we performed reliability and validity analyses (Studies 5 and 6).

A first approach to the reliability of the instrument was carried out by computing the Alpha coefficient in a sample of *n*=591 students from 4 different Universities, that gave a moderate value (Alpha=0.797). This value is reasonable, given that the questionnaire tries to assess a wide range of knowledge (see Table 1), so that a particular student can understand some of these concepts and do not understand others (Thorndike, 1991; Melia, 2001). We also computed two reliability coefficient based on factor analysis (Barbero, 2003):

1. $\theta = \frac{n}{n-1}\left(1-\frac{1}{\lambda_1}\right)$=0.82; was high, since the first eigenvalue explained a relatively high percentage of variance and most items contributed to that factor before rotation, which is an indication of an underlying construct being measured by the questionnaire.
2. $\Omega = 1-\frac{n-\sum h_j^2}{n+2\sum r_{jh}} = 0.896$**;** was still higher, according what is theoretically expected; this coefficient measures the commonalities (common factors) in the items.

In the same sample (*n=591*) we also carried out a generalizability analysis (López Feal, 1987; Feldt & Brennan, 1991; Martínez Arias, 1995), an approach that considers the different sources of error in measurement, analyses the component of these errors and provides different coefficients. In this method it is possible to fix some sources of errors and use the analysis of variance to estimate the different components in the total variance, including the variance of errors. We took into account two different sources of variations in the tests scores:

1. Generalizability of results to other items (fixing the students and considering the items as the only source of variation). We obtained a coefficient G_i=0.799; very close to the Crombach's Alpha value, as, in this case the generalizability coefficient coincides with Alpha; the small difference is due to round-off in the computations.

2. Generalizability of results to other students (fixing the items and considering the students as the only source of variation). We obtained a coefficient G_i=0.987, which indicates a very high possibility of extending the results to other students similar to those taking part in the sample, when the items are fixed.

Another estimation of reliability using test- retest was carried out in a sample of 106 students, each of which completed the questionnaire in two different occasions with about a month between the two applications. We obtained a test- retest reliability coefficient of 0.871 (Pearson correlation) and 0.861 (Spearman Rho), which are quite high. The Pearson's correlations coefficients between responses to same items in the two applications were all statistically significant and positive, ranging between 0.29 and 0.79. Split-half reliability coefficient (when considering each application as half the total questionnaire) gave very high values (0.91); the means, variances, inter-element covariances and correlations were very similar in the two occasions; all of which assures a high test-retest reliability. The computation of test-restest reliability was complemented with the estimation of confidence and credibility intervals for all the correlations coefficients.

CPR validity

We carried out different studies to provide evidences for the validity of the questionnaire that was considered a unitary construct according Messick (1989; 1995; 1998) and AERA/ APA/ NCME (1999):

1. The theoretical analysis of the questionnaire content as well as the results from experts' judgment served to justify *content validity*, by comparing the content evaluated by each item to the semantic units included in the semantic definition (Study 3).
2. Studying the questionnaire capacity to discriminate between two groups of psychology students before and after studying conditional probability served to justify *criteria validity* (Study 6.1). We used discriminant analysis (Cuadras, 1981; Afifi y Clark, 1990) to compare results from 208 students without instructions and 177 students with instructions. Most items discriminated between the two groups (significant difference); the scarce exceptions were items measuring psychological biases. The canonical correlation was equal to 0.697 and the probability of correct classification was 82.34%, all of which suggest good criteria validity for the questionnaire. This study was complemented with statistical summaries, difference tests, confidence and credibility intervals for the mean of the total scores in the two groups that again favoured the group with instruction.

3. We analysed the structure of responses to the questionnaire in a sample of n=591 students and compared with the assumed structure of the construct (Study 6.2) to study the *construct validity* (Muñiz, 1994; Martínez Arias, 1995). We performed an exploratory factor analysis (Tabachnick & Fidell, 2001). We expected the analysis confirm a main underlying construct, but, at the same time we also expected to find other factors that included the biases described in the literature and that would not correlate with the mathematical problem solving competence of students. All of this was confirmed in the Factor analysis (main components extraction; varimax rotation), which lead to two different groups of interrelated factors, as described in Section 6.1.

Details and statistical results of all the different steps in the process of building the questionnaire are included in Chapters 4 and 5 of the thesis. We applied Bayesian methods along all these steps, in order to fulfil the research objective 2. The result is the CPR questionnaire with reasonable reliability and validity that will be used in the next stage of the research and is also useful to other teachers and researchers.

6. DESING AND VALIDATION OF DIDACTIC RESOURCES TO INTRODUCE ELEMENTARY BAYESIAN INFERENCE IN PSICHOLOGY

Objective 3 in this research was *assessing conditional probability reasoning in psychology students to decide the suitability of teaching Bayesian methods to these students.* In Study 7 we applied the CPR questionnaire to a sample of 413 psychology students and analysed their responses from different points of views. Students showed enough understanding of conditional probability to start the learning of Bayesian inference, but, at the same time, we found some widespread misconceptions that were taken into account in the next stage (designing a curricular proposal).

Objective 4 in this research was *preparing and assessing didactic resources to introduce elementary Bayesian inference to Psychology students that takes into account the previous assessment.* To attain this aim we designed some teaching materials that were based on results of Study 7, some didactic principles and literature on teaching Bayesian inference. These materials were tried in Studies 8 and 9. Below we summarise these three studies, which are described in detail in Chapter 6 of the thesis.

6.1. ASSESSING CONDITIONAL REASONING IN PSYCHOLOGY STUDENTS

Once the CPR questionnaire was finished, we carried out an assessment study (Study 7).

Students from the Universities of Granada (4 different groups of students; *n*=308 students) and Murcia (two different group of students; *n*=106 students) took part in the sample (*n*=414). The students were enrolled in an introductory statistics course in the first year of University studies (typically, 18-19 year-olds). They had studied conditional probability at secondary school level and were taught conditional probability and the Bayes theorem with the help of tree diagrams, two-way tables and examples in the field of psychology, for about 2 weeks before they completed the questionnaire. The questionnaire was given to the students as an activity in the course of data analysis. Participation was optional and all the students were collaborative with the research.

Once the data were collected, we analysed the response of each student in each item. The scoring for open-ended items took into account the completeness of response. In items 2, 8, 11, 12, 13, 15 the students were given a point in case they identified correctly the problem data; correct built a tree diagram and identified the conditional probability, and 2 points for a totally correct solution. In item 1 and 16 the scoring ranged from 1 to 4 (see Table 4). The maximum possible scoring in the questionnaire was 34 points. The empiric distribution of scoring ranged between 3 and 30 with an average value of 19.12, a little higher than half the maximum possible score and the standard deviation was 5.91

In computing several probabilities from a two-way table (item 1) 90% of the students correctly computed the simple probability, 61%, the joint probability and 59% and 56%, respectively the two conditional probabilities. This confirms Falk´s (1989) opinion that verbal ambiguity in linguistic expression of conditional probability still makes it difficult for the student to distinguish conditional and joint probabilities after instruction.

Results in Table 2 suggest the existence of the following reasoning conflicts among the students in the sample:

Table 2. Percentage of responses in multiple-choice items (n=414)

	a	b	c	d	Blank
I3	8	7	29	50(+)	5
I4	28	15	29	20(+)	8
I5	1	89(+)	10	(*)	0
I7	35(+)	31	34	(*)	0
I9	25(+)	9	62	(*)	4
I10	6	32(+)	59	(*)	9
I14	77	9	10 (+)	2	2
I17a	6	17	69 (+)	7	1
I17b	24(+)	25	9	36	6
I18	9	13	76(+)	(*)	2

(*) Does not apply (+) Correct

1. *As regards independence*: we found confusion of independence with mutual exclusiveness in 28 % of the responses to distractor a) in item 4; a bias also noticed by Sánchez (1996). The chronological conception of independence described by Gras and Totohasina (1995) was also shown in 29% of the responses to distractor b) in item 4.
2. *Concerning conditional probability:* 31% of the students confused it with a joint probability (response b in item 7) or with a simple probability (34% responses c in item 7). The conjunction fallacy was observed in 62% of the responses to item 9 and the confusion of the transposed conditional in 59% of the responses in item 10. Difficulties in computing probabilities when the time axis is inverted are suggested by the responses to items 14 and 17b, although the chronological conception of conditional probability described by Gras and Totohasina (1995) was not so clearly shown in these two items.
3. The base rate fallacy was not as pervasive as suggested in previous research (Bar-Hillel, 1983) as shown in the responses to distractors (a) and (b) in item 3; since the majority of students gave the correct response (d) in this item, then showing improvement of base rate with instruction. Item 18 was also very easy.

Table 3. Completeness of solutions in open-ended items

	I1	I2	I8	I11	I12	I13	I15
Blank or totally wrong	29	15	47	18	21	30	24
Partly correct	30	21	18	21	9	18	16
Correct solution	41	64	35	61	70	52	60

As regards responses in open-ended items, results in Table 3 suggest that students had difficulties in giving a sound definition and an example of conditional probability (item 1) but were conscious of the restriction of sample space (item 2). They had difficulties in solving a conditional probability problem in a single experiment (item 8) due to a lack of distinction of dependent and independent experiments in the context (synchronic situation), so that many of them did not appear to have completely reached Level 4 in the conditional probability reasoning scheme by Tarr and Jones (1997).

Solving total probability (item 15) and solving conditional probability problems with replacement problems (item 12) and computing compound probability in the case of independent (item 11) events were easier than computing compound probability in dependent (item 13) events.

Table 4. Completeness of solutions in solving a Bayes problems (Item 16)

	Percentage
Blank or totally wrong	16
Correct identification of data	15
Identifies the inverse conditional probability,	16
Correct computation of denominator (total probability)	7
Correct solution	46

As regards solving an open Bayes problem (item 16), more than half the students were able to compute the total probability and a little less gave the complete solution; the majority was at least capable of correctly identifying the data and even identifying the probability to be computed although 16% failed in developing the total probability formula. We remark that data were given in the percentage format, which is considered harder than absolute frequency formats in Gigerenzer (1994) and Gigerenzer and Hoffrage´s (1995) research. We can conclude that, in general, the instruction was successful as regards problem solving capabilities, whenever there were no psychological biases involved in the situation. However, part of the biases described in the literature seemed not to be overcome with instruction.

To explore our conjecture that biases on conditional probability reasoning are unrelated to mathematical performance in the tasks, we carried out a factor analysis of the set of responses to the items (correct-incorrect responses to each item by the different students) using the SPSS software. The factor extraction method was principal components, which is the most conservative method, as it does not distort the data structure. In Table 5 we present the factor loadings (correlations) of items with the different factors after Varimax rotation (orthogonal rotation; maximizing variance of the original variable space). We found 7 factors with eigenvalue higher than 1 that explained the following percentages of the total variance: 21% (first factor), 7 % (second factor), and about 6% in the remaining factors; that is, a total of 59% of the variance was explained by the set of factors, which suggests the specificity of each item, and the multidimensional character of the construct, even when there is a common part shared by all of the items.

These percentages of variance also revealed the greater importance of the first factor, to which most of the open- ended problems contribute, in particular solving Bayes' problems had the higher contribution, followed by solving total probability and compound probability problems. All of these problems require a solving process with at least two stages, in the first of which a conditional probability is computed, which is used in subsequent steps (e.g. product rule). We could interpret this factor as *solving complex conditional probability*

problems ability.

Table 5. Factor Loadings for Rotated Components in Exploratory Factor Analysis of Responses to Items

Item	Component						
	1	2	3	4	5	6	7
Item 16. Bayes rule	.76						
Item 11. Total probability	.76						
Item 15. Product rule in dependent, synchronic events	.75						
Item 13. Product rule in independent events	.67						
Item 12. Conditional probability with replacement	.43		.42				
Item 6b. Conditional probability. Table		.79					
Item 6c. Joint probability. Table		.77					
Item 6a. Simple probability. Table	.32	.61					
Item 6d. Conditional probability. Table		.61					
Item 8. Conditional probability in single experiment			.67				
Item 1. Definition			.59				
Item 2. Sample space	.40		.45				
Item 17b. Time axis fallacy, diachronic experiment				.71			
Item 14. Time axis fallacy, diachronic experiment				.70			
Item 7. Cond prob. from joint and compound probability, synchronic					.66		
Item 9. Conjunction fallacy					.62		
Item 5. Conditional probability, without replacement, diachronic			.39		.44		
Item 17a. Conditional probability, without replacement						.66	
Item 10. Transposed conditional /causal-diagnostic						-.65	
Item 3. Independence /mutually exclusiveness							.68
Item 3. Base rates/ Bayes rule	.34						.48
Item 18. Product rule dependence, diachronic				.35			-.46

Computing simple, joint and conditional probability from a two-way table (item 6) appeared as a separate component, probably because the task format affected performance, a fact which has also been noticed by Ojeda (1996) and Gigerenzer (1994), among other researchers. A third factor showed the relationships between *definition, sample space and computation of conditional probabilities in, with and without replacement situations*; that is, we interpreted this factor as *Level 4 reasoning* in Tarr and Jones (1997) classification.

The remaining factors suggested that the different biases affecting conditional probability reasoning that are described in the justification, appeared unrelated to mathematical performance in problem solving, understanding, building the sample space and computing conditional probability, and to Tarr and Jones's (1997) level 4 reasoning (as related items were not included in the three first factors). Each of the biases (transposed conditional, time axis fallacy, conjunction fallacy, independence/mutually exclusiveness/synchronic setting) also appeared unrelated to one another; in some cases some of them were opposed or related to some semantic units in the mathematical component of understanding conditional probability. For example, independence was linked to the base rate fallacy (where people have to judge whether if the events are independent or not) and opposed to the idea of dependence.

In summary, these results supported our previous hypotheses that biases in reasoning about conditional probability are unrelated to mathematical performance in problem solving and, at the same time, support construct validity evidence for the questionnaire. At the same time it provides information about potential biases students might hold that were used in the design of the teaching experience in the next step of this research.

6.2. EVALUATION OF A TEACHING EXPERIENCE

There is nowadays a tendency to recommend that teaching of Bayesian inference might be included in undergraduate statistics courses as an adequate and desirable complement to classical inference (Lecoutre, 1999; 2006; Lecoutre, Lecoutre & Poitevineau, 2001; Iglesias, Leiter, Mendoza, Salinas & Varela, 2005). Situations where available a priori information can help making an accurate decision and software that facilitates the application of these methods are becoming increasingly available,

Some excellent textbooks whose understanding does not involve advance mathematical knowledge and where basic elements of Bayesian inference are contextualized in interesting examples (e.g., Berry, 1995 or Albert & Rossman, 2001) can help follow these recommendations. There are also a great number of Internet didactic resources that might facilitate the teaching of these concepts (e.g. those available from Jim Albert's web page at, http://bayes.bgsu.edu/). These and other authors (Bolstad, 2002) have incorporated Bayesian methods to their teaching and are suggesting that Bayesian inference is easier to understand than classical inference. This is however a controversial question (see Moore, 1997) and moreover empirical research that analyze the learning of students in natural teaching contexts is still very scarce.

The aim of Study 8 in this thesis was *to explore the possibility of introduce basic ideas of Bayesian inference to undergraduate psychology students and report the extent to which the learning goals were achieved.* The goal of Study 9 *was identifying groups of related concepts, as well as implications between learning objectives with the aim of providing some recommendations about how best organised the teaching of the topics*. In both studies we took into account the results of the previous assessment Study 7.

The sample taking part in this research included 78 students (18-20 year-olds) in the first year of the Psychology Major at the University of Granada, Spain. These students were in the introductory statistics course and volunteered to take part in the experiment. The sample was composed by 17.9% boys and 82.1% girls, which is the normal proportion of boys and girls in the Faculty. These students scored an average of 4.83 (in a scale 0-10) in the statistics course

final examination with standard deviation of 2.07.

The students were organized into four groups of about 15-20 students each and attended a short 12 hours long course given by the same lecturer with the same material. The 12 hours were organized into 4 days. Each day there were two teaching sessions with a half-an hour break in between. The first session (2 hours) was devoted to presentation of the materials and examples, followed by a short series of multiple choice items that each student should complete, in order to reinforce their understanding of the theoretical content of the lesson.

In the second session, students in pairs worked in the computer lab with some Excel programs provided by the lecturer to solve a set of inference problems. The Excel programs were as follows:

1. Program Bayes: This program computes posterior probabilities from prior probabilities and likelihood (that should be identified by the students from the problem statement).
2. The program Prodist transforms a prior distribution $P(p=p_0)$ for a population proportion p in the posterior distribution $P(p=p_0|data)$, once the number of successes and failures in the sample are given. Prior and posterior distribution is represented graphically.
3. The program Beta computes probabilities and critical values in the Beta distribution $B(s,f)$, where s and f are the successes and failures in the sample.
4. The program Mean computes the mean and standard deviation in the posterior distribution for the mean of a normal population, when the mean and standard deviation are given in the sample and prior population.

In table 1 we present a summary of the teaching content. Students were given a printed version of the didactic material that covered this content. Each lesson was organized in the following sections: a) Introduction, describing the lesson goals and introducing a real life situation; b) theory development, using the situation previously presented; c) additional examples of other situations where the same procedures and concepts could be applied, d) some solved exercises, with description of main steps in solving the exercises; e) new problems for students to solve in the computer lab; and f) self assessment. All this material together with the Excel programs was also made available to the students on the web site (http://www.ugr.es/~mcdiaz/bayes) and is also included as Appendix 7 to 9 in the thesis. We added a forum, so that students could consult the teacher or discuss themselves their difficulties, if needed.

Table 6. Teaching content and its organization

Lesson	Content	In classroom Session 1	Computer lab Session 2
1	Bayes theorem in the context of clinical diagnose	Prior and posterior probabilities; likelihood; Bayes theorem Subjective probability. Comparison with classical and frequentist probability. Revision of beliefs; sequential application of Bayesian procedures	Solving Bayes Problems: (Program Bayes)
2	Inference for proportion. Discrete case in the context of voting	Sample and population; parameters and statistics; Parameter as random variable; Prior and posterior distribution; Informative and non informative prior distribution. Credible intervals	Computing credible intervals for proportion; assigning non informative and informative prior distributions (Program Prodist)
3	Inference for proportion. Continuous case in the context of production	Generalization to continuous case. Beta distribution, its parameters and shape. Credible intervals; Bayesian tests	Assigning non informative and informative prior distributions Computing credible intervals for proportion; testing simple hypotheses (Program Beta)
4	Inference for the mean of a normal population in the context of psychological assessment	Normal distribution and its parameters; credible intervals and tests for the mean of a normal distribution with known variance; non informative and informative prior distributions	Assigning non informative and informative prior distributions Computing credible intervals for means; testing simple hypotheses (Program Mean)

Two weeks after the end of the teaching, the students were given a questionnaire to assess their understanding of the topic. They were warned to study the topic and prepare for the assessment and were motivated to get a good result in the test.

Questionnaire. A-priori analysis

The BIL (Bayesian Inference Learning) questionnaire (which is presented in Appendix) was made of multiple choice and some open ended items that were developed by the author with the specific aim to cover the most important contents in the teaching. In table 7 we describe the contents assessed by the different items in the BIL questionnaire. (In item *I*18 we considered three different scores). The aim was to assess learning in the following groups of concepts, which in our a-priori analysis were assumed to be the core content of basic Bayesian inference and might cause different types of difficulties to students. We also assumed learning of one of these groups of concepts would not automatically assure the learning of the other groups:

1. *Conditional probability and the Bayes' theorem.* As was argued before, different authors pointed to students' difficulties in understanding conditional probability: *fallacy of the transposed conditional*; causal and chronological conception of conditional probability; confusion between simple, joint and conditional probability. All these errors might cause difficulties in computing different types of probabilities (item2), understanding of the

differences between prior and posterior probability and likelihood (items 1 and 18), and using the Bayes' theorem as a tool to transform prior into posterior probabilities (item 7 and 18).

Table 7. Contents assessed in the BLI Questionnaire

Item	Content assessed
I1	Likelihood, conditional probability
I2a	Simple probability
I2b	Conditional probability
I2c	Conditional probability of contrary event
I2d	Joint probability
I3	Parameter as random variable
I4	Prior distribution
I5	Parameter as random variable; difference with statistics
I6	Correct assignment of a non informative prior distribution for proportion
I7	Using the Bayes' theorem as a tool to transform prior into posterior probabilities; table given
I8	Parameters in Beta distribution, defining prior informative distribution for proportion
I9	Parameters in Beta distribution,
I10	Computing credible intervals for proportion; reading Beta tables
I11	Testing simple hypotheses for proportion; reading Beta tables
I12	Properties of credible intervals
I13	Posterior distribution of mean; non informative prior. Known variance
I14	Testing simple hypotheses for means
I15	Posterior distribution of mean; non informative prior. unknown variance
I16	Credible intervals for means
I17	Posterior distribution for mean, informative prior
I18.1	Identifying prior probabilities from a problem statement
I18.2	Identifying likelihood from a problem statement
I18.3	Using the Bayes' theorem as a tool to transform prior into posterior probabilities;
I19	Meaning of likelihood
I20a	Parameters in Beta curve. Spread
I20b	Parameters in Beta curve. Centre

2. *Parameters as random variables, their distribution, distinction between prior and posterior distribution.* In Bayesian inference, parameters are considered to be random variables with a prior distribution, while in frequentist inference they are assumed to be unknown constants (items 3, 5), a distinction which is not too clear for some students (Bolstad, 2002). Moreover, the aim of Bayesian inference is to transform the prior into a posterior distribution via the Bayes' theorem (item 18). A prior distribution provides all the information for the parameter before collecting the data (item 4), non informative priors are given by uniform distributions and are used when no previous information is available for the parameter (item 6).

   There are different models to represent prior distributions. The Beta distribution was introduced in the teaching, and students had to learn the meaning of its parameters (item 8, 20) and how to select a specific Beta distribution in a particular inference problem (item 9). Students knew the normal distribution from previous lessons. However, they had to

learn the rule to compute the posterior distribution for a mean when the prior distribution is normal (item 13; 14, 15, 16). In managing all these distributions, Bayesian statistics uses the rules of probability to make inferences, and that requires dealing with formulae, but actual calculus used is minimal as students only have to understand that probability is given by different types of areas under a density function (Bosltad, 2002). However, the extent to which all of this is grasped by psychology students has still to be assessed.

3. *Logic of Bayesian inference.* The aim of Bayesian inference is updating the prior distribution via the likelihood to get the posterior distribution, which provides all the information for the parameter, once the data have been collected (Bolstad, 2004). However, it is also possible to carry out procedures similar to those used in frequentist statistics, although the interpretation and logic is a little different (Berry, 1995; Lecoutre, 2006). Credible intervals provide the epistemic probability that the parameter is included in a specific interval of values, for the particular sample, while confidence intervals provide the frequentist probability that in a percentage of samples from the same population the parameter will be included in intervals of values computed in those samples. Credible intervals are computed from the posterior distribution (item 17) and students should be able to compute them by using the tables of different distributions (items 10, 16); they should understand that the interval width increases with the credibility coefficient and decreases with the sample size (item 12).

   In Bayesian inference we can compare at the same time different hypotheses; in this case we compute the probabilities for those hypotheses given the data by using the posterior distribution and select the hypothesis with higher probability (item 11). In testing only one hypothesis we either compute the probability for the hypothesis or for the contrary event (item 14); acceptance or rejection will depend on the value of that probability. So, there are some conceptual and interpretative differences between classical and frequentist approaches, but, since both approaches often lead to approximately the same numerical results, students might not understand these differences and confuse both approaches (Iversen 1998)

Results

There were only 4 difficult tasks (percentage of correct responses under 50%). These tasks were (See table 8) the following: In item 14 (testing hypothesis about the mean) students either made an error in the reasoning by contradiction (choosing distractor c) or did not understand the standardization operation and choose distractor a). Of course this is a highly

complex item, where the logic of testing hypotheses is mixed with knowledge of probability calculus and standard Normal distribution. Students also found much difficulty in items 2b, and 2c where they confused a conditional probability and its inverse, a problem that have been repeatedly denounced (Bar-Hillel & Falk, 1982; Falk, 1986). We remark that distractors in this item are given only by formulas (instead of using a verbal description such as in item 1) while we found a high percentage of correct responses in item 1 and 7, in spite of the many difficulties and misconceptions described for conditional probability (see Batanero & Sánchez, 2005 for a survey). We conclude that the expressions prior and posterior probabilities and likelihood helped students to better distinguish a conditional probability and its inverse in these items. Finding a posterior distribution for the mean (item 15) was also difficult because students forgot to divide by the square root of the sample size to find the standard deviation in the posterior distribution. All the other tasks had a medium difficulty (between 50-60% correct responses).

Table 8. Results in BIL questionnaire

	% Correct responses	Confidence interval 95%		Credible interval 95%	
		Lim inf	Lim inf	Lim sup	Lim sup
1	88.7	0.808	0.966	0.784	0.943
2a	79.0	0.689	0.891	0.673	0.872
2b	38.7	0.266	0.508	0.276	0.511
2c	29.0	0.177	0.508	0.192	0.412
2d	51.6	0.392	0.639	0.394	0.635
3	66.1	0.543	0.779	0.537	0.766
4	58.1	0.458	0.779	0.456	0.695
5	61.3	0.492	0.734	0.488	0.723
6	50.0	0.376	0.624	0.366	0.604
7	93.5	0.874	0.996	0.845	0.973
8	53.2	0.408	0.656	0.409	0.650
9	85.5	0.767	0.943	0.746	0.921
10	64.5	0.526	0.764	0.520	0.752
11	58.1	0.458	0.704	0.456	0.695
12	53.2	0.408	0.656	0.409	0.650
13	69.4	0.579	0.809	0.570	0.793
14	30.6	0.191	0.421	0.206	0.429
15	40.3	0.281	0.525	0.290	0.527
16	69.4	0.579	0.809	0.570	0.793
17	69.4	0.579	0.809	0.570	0.793
18	79.0	0.689	0.891	0.673	0.872
19	58.1	0.458	0.704	0.456	0.695
20a	82.3	0.728	0.918	0.709	0.897
20b	72.6	0.615	0.837	0.582	0.800

Table 9. Results in problem solving in lesson 4 (Inference about a mean) (*n*=78)

		% Correct responses	95% Conf, interval		95% Credible interval	
			Lim inf	Lim inf	Lim sup	Lim sup
Ej.1	Correct solution	78.2	0.690	0.874	0.678	0.858
	Typify	83.3	0.750	0.916	0.724	0.891
	Identify the Z interval / Define hypothesis	84.6	0.766	0.926	0.750	0.909
	Compute final distribution	85.9	0.782	0.936	0.765	0.919
	Identify data	88.5	0.814	0.956	0.795	0.937
Ej.2	Correct solution	67.9	0.575	0.783	0.569	0.772
	Typify	87.1	0.797	0.945	0.780	0.928
	Identify the Z interval / Define hypothesis	88.5	0.814	0.956	0.795	0.937
	Compute final distribution	82.0	0.735	0.905	0.721	0.889
	Identify data	78.2	0.690	0.874	0.678	0.858

We also gave students problem solving activities and short self-assessment questionnaires in each lesson. In Table 9 we show results of solving problems related to inference in a mean (normal population). Details of results in the other intermediate assessment are included in Chapter 8 of the thesis and again show that students were capable of solve simple activities of Bayesian inference for proportions and means, including computing credible intervals and carrying out hypotheses tests.

6.3. INTERRELATIONSHIP BETWEEN CONDICIONAL PROBABILITY REASONING AND LEARNING OF BAYESIAN INFERENCE

To study the interrelations and implications between learning objectives we carried out several multivariate analyses, using the CHIC software, Classification Hierarchical, Implicative et Cohesive (Couturier and Gras, 2005). The implication index between two dichotomous variables *a* and *b* in a population is defined by

$$q(a,\overline{b}) = \left[\frac{card(A \cap \overline{B}) - \frac{card(A)card(\overline{B})}{n}}{\sqrt{\frac{card(A)card(\overline{B})}{n}}} \right]$$

where *A* and *B* are the population subgroups where *a* and *b* take the value 1 (Gras, 1993; 1996; Gras & Ratsima-Rajohn, 1996). This index follows the normal distribution *N(0,1)*, and from there an intensity for the implication $a \Rightarrow b$ is defined by

$$\varphi(a,\overline{b}) = \Pr ob[car(X \cap \overline{Y}) \le card(A \cap \overline{B})],$$

where *X* and $\overline{Y}$ are dichotomous independent random variables having the same cardinal than *A* and $\overline{B}$ respectively (Lerman, Gras & Rostam, 1981a & b). In our study we have a total of $C_{21,\,2}$ implication indexes among the 21 subítems in the LBI questionnaire. The software

CHIC computes these indexes and provides a graph with all the implications which are significant to a given significance level.

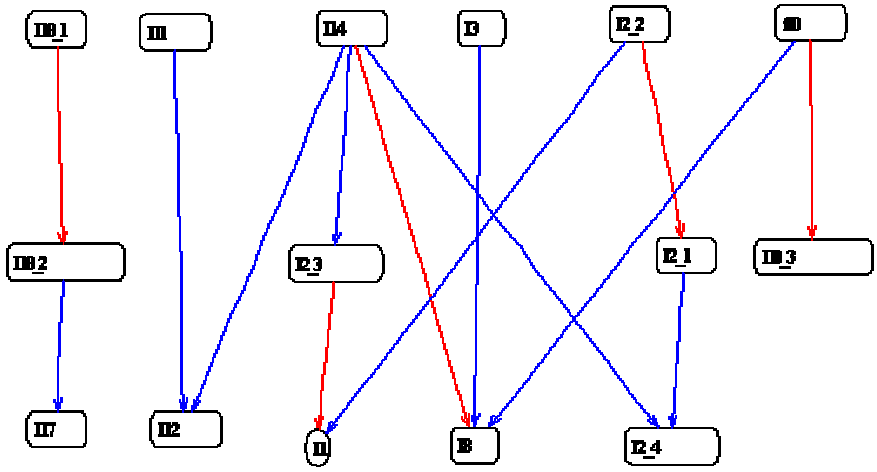
The implication $a \Rightarrow b$ in our study is interpreted in the sense that when a students correctly solves item *a* there is higher probability for him /her to solve item *b*. In this sense the implicative graph provides a possible order to introduce different concepts and procedures whose understanding is assessed in those items in the teaching of the topic. Before carrying out the implicative analysis we checked the assumptions of the method; experimental units of variables, and independence of responses by different students. We assumed a binomial model for the responses; that is, we assumed each student having same likelihood to correctly solve the items (Lerman, 1991), as in fact these are the hypotheses assumed in classical theory of tests.

In Figure 2 we present the implicative graph with all the relationship that were significant at 99% level (red) or 95% level (blue), We observe that the implication relationship is asymmetrical and the sense of implication is showed by the arrows in the graph.

If we study the relationships higher than 99% in the graph, we observe that students who correctly answer item I18_2 (correct identification of likelihood, which is given by a conditional probability) have better likelihood to answer I18_1 (correct identification of prior probabilities, which are given by simple probabilities). Correct performance in I10 (identifying probabilities and critical values from the Beta distribution table and computing credible intervals for a proportion) facilitate correct computation of posterior probabilities with Bayes theorem (I18_3). Both tasks involve computing probabilities but the first one is more complex. Then correct computation of conditional probabilities implies correct computation of join and single probabilities (I2_1, I2_2, I2_4).

As regards implications higher than 95% (blue in the diagram) we observe that students who correctly perform a Bayesian hypothesis test (I14 or I11) increase their likelihood to correctly interpret credible intervals (I12), possibly because all the ideas in understanding the second task are involved in the first one, which adds the need to understand the logic of proof by contradiction. I14 implies I2_3, the computation of conditional probability for a contrary event, but, again mastering the idea of proof by contradiction involves correct reasoning on both conditional reasoning and complementation. Students who visualize parameters as radon variables (I3) or compute probabilities for Beta function and credible intervals for proportions (I10) perform better in correctly assigning a Beta informative prior distribution (I8), a task that is also facilitated by I14.

Figure 2. Implicative graph with significant implications at 99 and 95%



I2_3 (computing the conditional probability for the contrary event) or I2_2 (computing conditional probability) facilitates I1, distinguishing prior and posterior probabilities and likelihood (all these ideas are supported on correct conditional reasoning); I2_2 facilities computing simple probability (I1) and both of them together facilitate the computation of join probabilities (I2_4), another task which is easier for those who succeeded in I14 (testing hypotheses).

Implicative hierarchy of learning outcomes

Once the isolate implications between items were studied we carried out an implicative classification analysis. This is an algorithm, which uses the implicative indexes in a set of variables to study the internal cohesion of some variables subsets (Lahanier-Reuter, 2001; Couturier, Gras & Guillet, 2004). The cohesion between two variables *a* and *b* is defined by $c(a,b)=\sqrt{1-H^2}$ where H is the entropy for the two variables, and varies between 0 and 1. The cohesion for a class of variables is defined by (Gras, Kuntz & Briand, 2001):

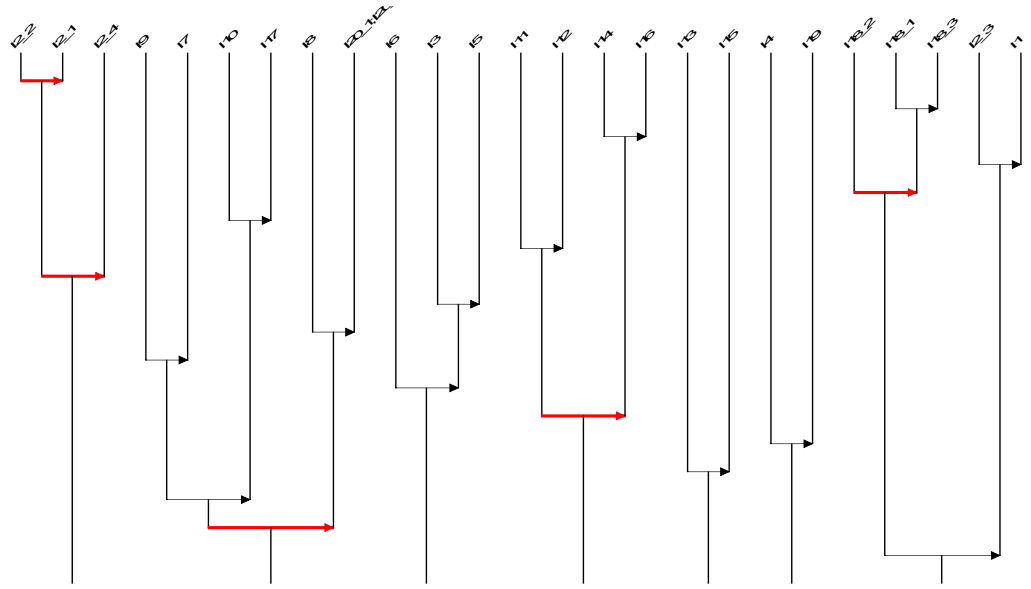
$$C(\underline{A})=\left[\prod_{j\in\{2,\ldots r\},\, j>i}^{i\in\{1,\ldots r-1\}} c(a_i,a_j)\right]^{\frac{2}{r(r-1)}}$$

Then, given two sets of variables $\underline{A}$ and $\underline{B}$ the strength of implication from $\underline{A}$ to $\underline{B}$ is defined by (Couturier, 2001):

$$\psi(\underline{A},\underline{B})=\left[\sup_{i\in\{1,\ldots r\}\, j\in\{1,\ldots s\}} \varphi(a_i,\bar{b}_j)\right]^{rs}\left[C(\underline{A}).C(\underline{B})\right]^{1/2}$$

The software CHIC builds an implicative hierarchy in the set of variables, taking into account both the maximal cohesion into each class and the higher implication from a class to another. In Figure 3 we present the hierarchy produced. There are four significant clusters:

Figure 3. Implicative hierarchy with 95% node



- Group 1. Items (I2_2) and (I2_1) which join to (I2_4), all of them related to probability. The student who correctly computes conditional probabilities (I2_2), correctly perform simple (I2_1) and compound probability (I2_4). The higher difficulty of conditional probability as regards simple and compound is then confirmed.
- Group 2: Prior and posterior distributions and Beta curves. Item I9, I7, I10, I17, I8 and the two parts of I20. Students who are able to interpret the parameters in the Beta curve (I9) and understand how posterior distributions are get from prior distributions and likelihood through Bayes theorem (I7) succeeded better in getting a credible interval for proportions in the continuous case; a task that requires interpreting probabilities of Beta curves, and understanding the concept of posterior probability, as well as the concept of credible interval. They also performed better in discriminating prior and posterior distribution of the mean (I17). All of this lead to better choosing a non informative prior distribution for proportion in the continuous case through the Beta Curve (I8) and graphically interpreting the parameters in Beta curves (I20).
- Group 3 (Items I11, I12, I14 and I 16) group a set of Bayesian inference tasks. Being able of correctly test a hypothesis for proportions (I11) increases the likelihood of correctly

interpret credible intervals (I12); and these two task are associated with correctly testing a hypothesis about the mean (I14), and correctly computing a credible interval for the mean (I16). All these are knowledge specifically related to the Bayesian methods which are based on conditional probability and also in the logic of scientific inference.

- Group 4: Moreover there is a second group of tasks related to conditional probability (the different parts of Item 18, I2-3 and I1). Correct identification of likelihood from a problem statement (I18_2) facilitates correct identification of prior probability (I18-1) and this lead to correct computation of posterior probabilities (I18_3). These three abilities lead to better identification of conditional probabilities for the contrary event (I2-3) and discrimination between prior probability, likelihood and posterior probabilities in the context of a problem (I1).

Other groupings of items that are non significant were as follows:

- Group 5: Items I6 (assigning adequate prior distribution for the non informative case to proportions in the discrete case), I3 (understanding parameters as random variables) and I5 (discrimination between parameters and statistics); all these tasks are related to understanding parameters from a Bayesian point of view.
- Group 6: Items I13 (Posterior distribution of mean when variance is known) and I15 (posterior distribution of mean when variance is unknown: related to specific knowledge the students should remember.
- Group 7: I4 (concept of prior distribution) and I19 (concept of likelihood).

In summary these implications point to three groups of concepts relevant for students' introduction to the elementary ideas of Bayesian inference and that should be taken into account in planning the teaching and support our previous a-priori analysis of the BIL questionnaire:

1. *Conditional probabilistic reasoning* (as shown in groups 1 and 4), a theme where many biases have been described in the literature, but which is basic in defining posterior probabilities and distributions and likelihood, as well as in understanding the logic of credible intervals and hypothesis testing. Results also suggested that formulas for different types of probability were harder than verbal expressions for students to understand. Perhaps we should take into account Feller's suggestion (1973, p. 114) that "*conditional probability*

is a basic tool of probability theory, and it is unfortunate that its great simplicity is somewhat obscured by a singularly clumsy terminology".

2. *Probability distributions,* its parameters (visualized as random variables), the distinction of prior and posterior distribution of parameters and assignment of prior distributions for informative and non informative cases (Groups 2, 5, 6 and 7). In our teaching we limited to Beta and Normal distributions, since the time available for teaching was restricted, but still so, the understanding of Beta curves appeared as a separated subgroup, as well as remembering the rules for known and unknown variance in inference about normal distributions. The difficulties to understand the different conception of parameters in Bayesian and frequentists statistics also appeared as a separated subgroup.
3. *Logic of Bayesian inference* (Group 3), that is, understanding the logic for computing and interpreting credible intervals and testing simple hypothesis. Performance in these tasks is in fact supported in understanding the previous two groups of concepts, most of which are not specific to Bayesian reasoning. However, limitation of teaching time leads some lecturers to reduce the teaching of the same and to try to pass directly from data analysis to inference. Teaching of Bayesian inference therefore should only be started when previous groups of concepts are well understood by students.

7. SUMMARY AND MAIN CONTRIBUTIONS

In this Thesis we focus on the use of Bayesian inference in the field of Psychology from different perspectives. Below we summarise these perspectives and the main conclusions /contributions achieve for each of them.

Current practice of statistics

We produced a synthesis of main criticisms of current statistical practices in psychology, the reported errors and the possible contribution of Bayesian inference to solve part of the denounced errors. As a consequence we suggested the need to introduce the teaching of elementary Bayesian methods in psychology and to carry out empirical research to assess the suitability of this teaching.

Application of Bayesian methods in psychometrics

We analysed the implications of a Bayesian approach to Classical Tests Theory and deduced estimation procedures for some of the psychometrics features of items and questionnaire. These procedures were applied in the process of building a questionnaire to

assess conditional probability reasoning (CPR), which is also justified in the thesis. We also developed some Excel programmes to carry out the main computations.

Assessing conditional probability reasoning in undergraduates

We used the CPR questionnaire to carry out a detailed evaluation in a sample of 414 students after teaching of the topics. The complex relationship between probabilistic concepts and intuition was shown in our study, where probabilistic biases were widespread in students, even in those with good problem solving probability. Consequently, our research suggest the need of reinforcing the study of conditional probability in the teaching of data analysis at University level, although it also provides arguments for a change of approach in this teaching. Following Nisbett and Ross' recommendations (1980, p. 280) students should be "*given greater motivation to attend closely to the nature of the inferential tasks that they perform and the quality of their performance"* and consequently "*statistics should be taught in conjunction with material on intuitive strategies and inferential errors*" (p. 281) of the kind presented in their book. In this sense we support Rossman and Short (1995), who suggest conditional probability can be taught in line with new statistics education ideas, in presenting a variety of applications to realistic problems, proposing interactive activities and using technology to facilitate learning.

Studying the suitability of teaching elementary Bayesian inference to undergraduates

We developed a teaching material that takes into account the previous analyses, as well as previous research in statistics education and the type of students. This material was trailed with a sample of 78 students, and data on the students' learning at the end of the experience showed that most instructional objectives were achieved by the students.

The implicative and cohesive classification analyses also supported the interrelationship between learning Bayesian inference and understanding conditional probability as it was previously assumed. On the other hand, the obtained classes in the implicative hierarchy provided us with information about the concepts whose understanding is related and their relative difficulty. This is a potential help to prepare didactic materials and to organize the teaching of the topic.

In summary, we think that this thesis opens a new perspective for research in the Behavioural Sciences Research Methods, both from the strictly methodological point of view (implementing and applying Research Methods) and from the didactic point of view. Partial

results of each of the mentioned contributions have been published in diverse journals and international conferences (See appendix 3).

In the present convergence process to the European Space of Higher Education, it is not only possible, but required that lecturers in this area carry out research on the didactics of research methods, including non-traditional topics. Only by means of systematic research we can enrich our educational practice and contribute to improve the application of research methods. It is therefore expected that new studies continue the research started in this Thesis.

APPENDIX 1. CPR QUESTIONNAIRE

Item 1. Explain in your own words what a simple and a conditional probability is and provide an example for each.

Item 2. Complete the sample space in the following random experiments:
a) Observing gender (male/female) of the children in a three children family (e.g. MFM,...)
b) Observing gender (male/female) of the children in a three children family when two or more children are male.

Item 3 (Tversky & Kahneman, 1982a)
A witness sees a crime involving a taxi in a city. The witness says that the taxi is blue. It is known from previous research that witnesses are correct 80% of the time when making such statements. The police also know than 15% of the taxis in the city are blue, the other 85% being green. What is the probability that a blue taxi was involved in the crime?

a. 80/100
b. b) 15 /100
c. (15/100) X (80/100)
d. $\frac{15 \times 80}{85 \times 20 + 15 \times 80}$

Item 4. (Sánchez, 1996)
A standard deck of playing cards has 52 cards. There are four suits (clubs, diamonds, hearts, and spades), each of which has thirteen numbered cards (2,..., 9, 10, Jack, Queen, King, Ace). We pick a card up at random. Let A be the event "getting diamonds" and B the event ·getting a Queen". Are events A and B independent?

a. They are not independent, since there is the Queen of diamonds
b. Only when we first get a card to see if it is a diamond, return the card to the pack and then get a second card to see if it is a Queen.
c. They are independent, since P(Queen of diamonds)= P(Queen)) x P(diamonds),
d. They are not independent, since P(Queen /diamonds) $\neq$ P(Queen).

Item 5.
There are four lamps in a box, two of which are defective. We pick up two lamps at random from the box, one after another, without replacement. Given that the first lamp was defective:

a. The second lamp is more likely to be defective
b. The second lamp is most likely to be correct.
c. The probabilities for the second lamp being either correct or defective are the same.

Item 6. (Estepa, 1994)
In a medical centre a group of people were interviewed with the following results:

	55 years-old or younger	Older than 55	Total
Previous heart stroke	29	75	104
No previous heart stroke	401	275	676
Total	430	350	780

Suppose we select at random a person from this group:
a. What is the probability that the person had a heart stroke?
b. What is the probability that the person had a heart stroke and, at the same time is older than 55?
c. When the person is older than 55, what is the probability of having had a heart stroke?
d. When the person had a heart stroke, what is the probability of being older than 55?

Item 7. Eddy (1982)
10.3 % of women in a given city have a positive mammogram. The probability that a woman in this city has both positive mammogram and a breast cancer is 0.8%- A mammogram given to a woman taken at random in this population was positive. What is the probability that she actually has breast cancer?
a. $\frac{0.8}{10.3} = 0.0776$, 7.76%
b. $10.3 \times 0.8 = 8.24$, 8.24%
c. 0.8 %

Item 8.
In throwing two dice the product of the two numbers was 12. What is the probability that none of the two numbers is a six (we differentiate the order of numbers in the two dice).

Item 9. (Tversky & Kahneman, 1982 b)
Suppose a tennis player goes to the Roland Garros posterior in 2005. He has to win 3 out of 5 sets to win. Which of the following events are more likely?
a. The player wins the first set
b. He wins the first set but looses the match
c. Both events a) and b) are equally likely

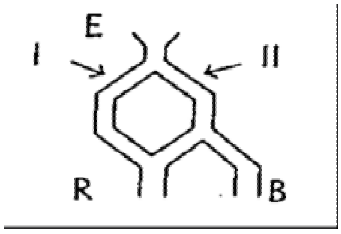
Item 10. (Pollatsek at al. 1987)
A cancer test was given to all the residents in a large city. A positive result was indicative of cancer and a negative result of no cancer. Which of the following results is more likely?
a. That a person had cancer if they got a positive result
b. Having a positive test if the person had cancer.
c. The two events are equally likely.

Item 11.
60% of the population in a city are men and 40% women. 50% of the men and 35% of the women smoke. If we pick a person from the city at random, what is the probability that the person is a smoker?

Item 12.
A person throws a die and writes down the result (odd or even). It is a fair die (that is all the numbers are equally likely). These are the results after 15 throws:
Odd, even, even, odd, odd, even, odd, odd, odd, odd, even, even, odd, odd, odd
The person throws once more. What is the probability to get an odd number this time?

Item 13.
A group of students in a school take a mathematics test and an English test.80% of the students pass the mathematics test and 70% of the students pass the English test. Assuming the two subjects score independently, what is the probability that a student passes both tests (mathematics and English)?

Item 14. Ojeda (1996)
We throw a ball in the entrance E of a machine (see the figure).
If the ball goes out through R, what is the probability of having passed by channel I?
a. 1/2
b. 1/3
c. 2/3
d. Cannot be computed



Item 15.
According to a recent survey, 91% of the population in a city usually lie and 36% of them usually lie about important matters. If we pick a person at random from this city, what is the probability that the person usually lies about important matters?

Item 16. Totohasina (1982)
Two machines M1 and M2 produce balls. Machine M1 produces 40 % and M2 60% of balls. 5% of the balls produced by M1 and 1% of those produced by M2 are defective. We take a ball at random and it is defective. What is the probability that that ball was produced by machine M1?

Item 17. (Falk, 1986, 1989)
Two black marbles and two white marbles are put in an urn. We pick a white marble from the urn. Then, without putting the white marble in the urn again, we pick a second marble at random from the urn.
1. If the first marble is white, what is the probability that this second marble is white? P (N_2/ N_1)
a. 1/ 2
b. 1/ 6
c. 1/ 3
d. 1/ 4

2. If the second marble is white, what is the probability that the first marble is white? P (N_1/ N_2)
a. 1/ 3
b. Cannot be computed
c. 1/ 6;
d. 1/ 2

Item 18.
An urn contains one blue marble and two red marbles. We pick up two marbles at random, one after the other without replacement. Which of the events below is more likely?
a. Getting two red marbles.
b. The first marble is red and the second is blue
c. The two events a) and b) are equally likely.

APPENDIX 2. LBI QUESTIONNAIRE

Item1. 10 out of every 100 students in a Faculty study mathematics; 30 out of every 100 students doing mathematics share an apartment with other students. Let *S* be the event "sharing the apartment" and M the event the student is doing mathematics course. If we pick a student at random and the student is doing mathematics, the probability that he shares the apartment is:
1. A prior probability *P (S)*
2. A posterior probability *P(S|M)*
3. A likelihood *P(M|S)*
4. A joint probability *P (M∩S)*

Item 2. Imagine you pick 1000 people at random. You know that 10 out of every 1000 people get depression. A depression test is positive for 99 out of every 100 depressed people as well as for 2 out of every 100 non depressed people. Given that D means depression and + means a positive test, compute the following probabilities:
1. *P(D)=* *2. P(+|D)=* *3. P(-|D)=* *4. P(D∩+)=*

3. The mean value μ for a variable (for example height) in a population:
1. Is a constant in Bayesian inference
2. Is a random variable in classical inference
3. **Is a random variable in Bayesian inference**
4. Could be constant or variable, depending on the population

Item 4. The prior probability distribution for a parameter:
1. **Provides all the information about the population before collecting the data**
2. Is computed from the posterior distribution by using the Bayes theorem.
3. It can be used to compute the credible interval for the parameter
4. Is an uniform distribution

Item 5. 1000 young Spanish people were interviewed in a survey. On average they spent 3 hours a week in practicing some sports. In Bayesian inference:
1. 3 hours is a parameter in the population of young Spanish people
2. **The average in this population is a random variable; the most likely value is about 3 hours.**
3. The average in this population is an unknown constant.
4. Each young Spanish person spends 3 hours a week in doing some sport.

Item 6. In a factory lamps are sold in boxes of four lamps. We have no information about the proportion of defective lamps. Which of the distributions A, B. C or D better describes the prior distribution for the proportion of defective lamps in a box?

(A)		**(B)**		(C)		(D)	
Values of Proportion	Probability	**Values of proportion**	**Probability**	Values of proportion	Probability	Values of Proportion	Probability
0.00	0.1	**0.00**	**0.2**	0.00	0.00	0.00	1/4
0.25	0.1	**0.25**	**0.2**	0.01	0.25	0.25	1/4
0.50	0.1	**0.50**	**0.2**	0.02	0.50	0.50	1/4
0.75	0.1	**0,75**	**0.2**	0.03	0.75	0,75	1/4
1	0.1	**1**	**0.2**	0.04	1	1	1/4

Item 7. In trying to estimate a proportion a student filled three columns in the Bayes table. He got these data:

Values of proportion	Prior Probability	Likelihood	--------	---------
0.0000	0.0000	0.0000		
0.1000	0.1000	0.0000		
0.2000	0.1000	0.0233		
0.3000	0.1000	0.1239		
0.4000	0.1000	0.0682		
0.5000	0.1000	0.0065		
0.6000	0.1000	0.0001		
0.7000	0.1000	0.0000		
0.8000	0.1000	0.0000		
0.9000	0.1000	0.0000		
1.0000	0.1000	0.0000		
Sum		0.0222		

The posterior probability that the true value of proportion in the population is 0.4 would be:
1. 0.00682 2. 0.1000 3. **0.3072** **4.** 0.00015

Item 8. A clinical survey showed a 15% incidence of tobacco addition in young women. A possible prior distribution to approximately describe this proportion is:
1. *B (15, 100)* 2. ***B (15, 85)*** ***3.*** *B (85,15)* *4. B (100, 15)*

Item 9. The mean for a *Beta B(a,b)* distribution is:
1. *a/b* *2. (a+1)/(a+b)* *3. (a+1)/(b+1)* *4.* ***a/(a+b)***

Item 10. In the following table probabilities and critical values for the B(30,40) distribution are given

Probabilities			Critical values	
p_0	$P(0<p<p_0)$	$P(p_0<p<1)$	$P(0<p<p_0)$	p_0
0	0.000	1.000	0.000	0.000
0.05	0.000	1.000	0.005	0.296
0.1	0.000	1.000	0.010	0.304
0.15	0.000	1.000	0.015	0.311
0.2	0.000	1.000	0.020	0.316
0.25	0.001	0.999	0.025	0.320
0.3	0.012	0.988	0.030	0.324
0.35	0.090	0.910	0.035	0.327
0.4	0.318	0.682	0.040	0.330
0.45	0.645	0.355	0.045	0.330
0.5	0.886	0.114	0.050	0.333
0.55	0.979	0.021	0.950	0.526
0.6	0.998	0.002	0.955	0.529
0.65	1.000	0.000	0.960	0.533
0.7	1.000	0.000	0.965	0.536
0.75	1.000	0.000	0.970	0.541
0.8	1.000	0.000	0.975	0.545
0.85	1.000	0.000	0.980	0.551
0.9	1.000	0.000	0.985	0.558
0.95	1.000	0.000	0.990	0.567
1	1.000	0.000	1.000	1.000

The 98 % credible interval for the proportion in a population described by a posterior distribution *B (30, 40)* is about:
1. $(0.316 < p < 0.551)$ 2. **$(0.304 < p < 0.567)$** **3.** $(0.3 < p < 0.6)$ 4. $(0.1 < p < 0.9)$

Item 11. The posterior distribution for the proportion of voters favorable to a political party is given by the *B (30, 40)* distribution. From the above data table, the most reasonable decision is accepting the following hypothesis for the population proportion
1. *H:* $p<0.25$
2. *H:* $p > 0.55$
3. ***H:* $p> 0.25$**
4. *H:* $p >0.45$

Item 12. For the same posterior distribution of the parameter in a population the r% credible interval for the parameter is:
1. **Wider if r increases**
2. Wider if the sample size increases
3. Narrower if r increases
4. It depends on the prior distribution

Item 13. In a normal population with standard deviation σ=5 and with no prior information about the population mean, we pick a random sample of 25 elements and get a sample mean $\bar{x} = 100$. The posterior distribution of the population mean is:
1. A normal distribution *N (100, 0,5)*
2. A normal distribution *N (0,1)*
3. A normal distribution *N (100,5)*
4. **A normal distribution *N (100, 1)***

Item 14. To test the hypothesis that the mean μ in a normal population with standard deviation σ=1 is larger than 5, we take a random sample of 100 elements. To follow the Bayesian method:

1. We compute the sample mean $\bar{x}$ and then compute $P(\frac{\bar{x}-5}{0,1} < 5)$; when this probability is very small, we accept the hypothesis.
2. **We compute the sample mean $\bar{x}$ and then compute $P(\frac{\bar{x}-5}{0,1} < Z)$; when Z is the normal distribution N (0,1); when this probability is very small, we accept the hypothesis.**
3. We compute the sample mean $\bar{x}$ and then compute $P(\frac{\bar{x}-5}{0,1} > Z)$ when *Z* is the normal distribution N (0,1); when this probability is very small, we accept the hypothesis.
4. We compute the sample mean $\bar{x}$ and then compute $P(\frac{\bar{x}-5}{0,1} > 5)$ when this probability is very small, we accept the hypothesis.

Item 15. In a sample of 100 elements from a normal population we got a mean equal to 50. If we assume a prior uniform distribution for the population mean, the posterior distribution for the population mean is:
1. About *N (50, s),* where s is the sample standard estimation.
2. **About *N (50, s/10),* where s is the sample standard estimation.**
3. We do not know, since we do not know the standard deviation in the population
4. About *N (0,1)*

Item 16. The posterior distribution for a population mean is *N (100, 15).* We also know that *P (-1.96 < Z < 1.96) =0.95*, where *Z* is the normal distribution *N (0,1).* The 95% credible interval for the population mean is:
1. (100-1.96 x 1.5; 100 + 1.96 x1.5)
2. (100-1.96; 100+1.96)
3. (100 x 1.5 –1.96; 100 x 1.5 + 1.96)
4. **(100-1.96 x 15; 100 +1.96 x 15)**

Item 17. In a survey to 100 Spanish girls the following data were obtained:

	Mean	Standard dev.
Sample	160	10
Prior distribution	156	13
Posterior distribution	158.5	7.9

To get the credible interval for the population mean we use:
1. The normal distribution *N(160,10)*
2. The normal distribution *N (156,13)*
3. **The normal distribution *N (158.5; 7.9)***
4. The normal distribution *N(160, 0.5)*

Item 18. 20 % of boys and 10% of girls in a kindergarten are immigrant. There are about 60% boys and 40% girls in the center. Use the following table to compute the probability that an immigrant child taken at random is a boy.

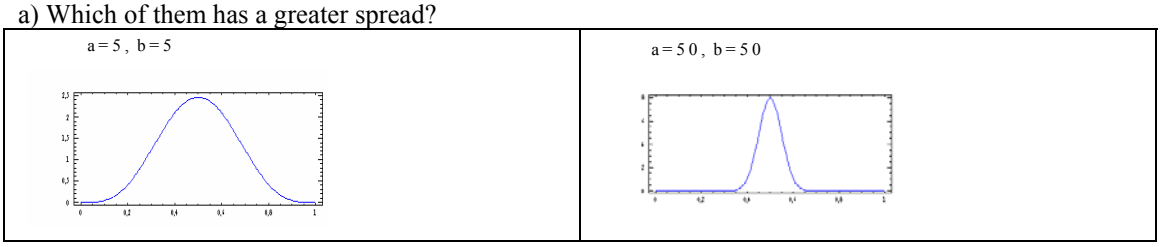
Events	Prior probabilities	Likelihoods	Product	Posterior probabilities
Sum	1		1	

Item 19. In a geriatric center we want to estimate the proportion of residents with cognitive impairment. 2 out of 10 residents taken at random in the residence showed cognitive impairment. The likelihood for the parameter p=0.1 is 0.1937. What is the meaning of this value?
1. *P (data),* that is, probability of getting this sample.
2. *P (data ∩ p=0.1),* that is, probability of getting the sample and that, in addition, the population proportion is 0,1.
3. *P (p=0.1|data),* that is, probability of a population proportion is 0.1. given the sample

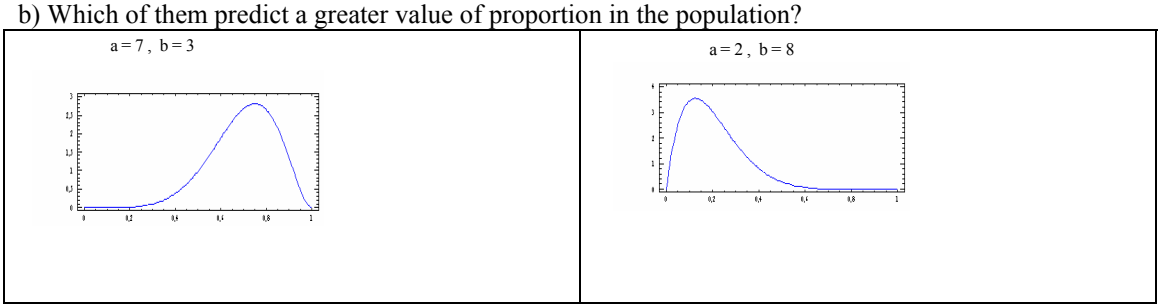
4. *P (data| p=0.1),* that is, given than p= 0.1, probability of getting this sample.

Item 20. Observe the following Beta curves

a) Which of them has a greater spread?



b) Which of them predict a greater value of proportion in the population?



APPENDIX 3. REFERENCES IN THE THESIS

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test? *Psychological Science, 8 (1)*, 12 – 14.

Aydýnlý, G., Härdle, W. y Rön, B. (2003). E-learning/e-teaching of statistics: A new challenge. En J. Engel (Ed.), *Proceedings of the IASE Satellite Conference Statistics Education and the Internet*. Berlin: International Association for Statistics Education. On line: http://www.stat.auckland.ac.nz/~iase/publications/6/Haerdle.pdf.

Agnoli, F. (1989). Suppressing natural heuristics by formal instruction: The case of the conjunction fallacy. *Cognitive Psychology*, *21(4),* 515.

Albert, J. (1995). Teaching inference about proportion. Using Bayes and discrete models. *Journal of Statistics Education, 3(3).* On line: http://www.amstat. org/publications/jse/v3n3/albert.html.

Albert, J. (1996). *Bayesian computation using Minitab*. Belmont, CA: Duxbury Press.

Albert, J. (2000). Using a sample survey project to assess the teaching of statistical inference. *Journal of Statistics Education, 8(1).* On line: http://www.amstat. org/publications/jse/secure/v8n1/albert.html.

Albert, J. (2002). Teaching introductory statistics from a bayesian perspective. En B. Philips (Ed), *Proceedings of the 6th International Conference on Teaching Statistics.* Ciudad del Cabo, Sudáfrica: International Statistical Institute.

Albert, J. H. y Rossman, A. (2001). *Workshop statistics. Discovery with data. A bayesian approach.* Bowling Green, OH: Key College Publishing.

Alvarado, H. (2004). *Significado del teorema central del límite en textos universitarios para ingenieros.* Trabajo de Investigación Tutelada. Universidad de Granada.

American Psychological Association (2001). *Publication manual*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ares, V. M. (1999). La prueba de significación de la «hipótesis cero» en las investigaciones por encuesta. *Metodología de Encuestas, 1*, 47-68.

Artigue, M. (1990). Epistémologie et didactique. *Recherches en Didactique des Mathématiques*, *10(2/3)*, 241–286.

Azorín, F. y Sánchez-Crespo, J. L. (1986). *Métodos y aplicaciones del muestreo.* Madrid: Alianza Editorial.

Ayçaguera, L y Benavides, A. (2003). Apuntes sobre subjetividad y estadística en la investigación en salud. *Revista Cubana de Salud Pública, 29(2),* 170-173. On line: http://scielo.sld.cu/scielo.php?pid=S0864-34662003000200012&script=sci_arttext& tlng=es.

Ayçaguera, L. y Suárez, P. (1995). ¿Qué es la inferencia bayesiana? *JANO,* 1132, 1542. On line: http://www.atheneum.doyma.es/Socios/sala_l/lec06est.htm.

Bailleul, M. y Gras, R. (1994). L'implication statistique entre variables modales. *Mathématiques, Informatique et Sciences humaines*, *128*, 41-57.

Bailleul, M. (2001). Des réseaux implicatifs pour mettre en évidence des representations. *Mathématiques et Sciences humaines*, *135,* 154-155.

Bakan, D. (1997). The test of significance in psychological research. En D. E. Morrison y R. E. Henkel, (Eds.). *The significance tests controversy: A reader* (pp. 231 – 251). Chicago: Aldine.

Bar – Hillel, M. (1983). The base rate fallacy controversy. En R. W. Scholz (Ed.), *Decision making under uncertainty.* (pp 39 – 61) Amsterdam: North Holland.

Bar – Hillel, M. (1987). The base rate fallacy controversy. En R. W. Scholz (Ed.), *Decision making under uncertainty.* (pp 39 – 61) Amsterdam: North Holland.

Barbero, M. (2003). *Psicometría II. Métodos de elaboración de escalas*. Madrid: UNED.

Batanero, C. (2000). Controversies around significance tests. *Mathematical Thinking and Learning, 2(1-2),* 75–98.

Batanero, C. (2001). (Ed.). *Training researchers in the use of statistics.* Granada: International Association for Statistical Education e International Statistical Institute.

Batanero, C. y Díaz, C. (2005). Análisis del proceso de construcción de un cuestionario sobre probabilidad condicional. Reflexiones desde el marco de la TFS. En A. Contreras (Ed.). *Investigación en Didáctica de las Matemáticas. I Congreso Internacional sobre Aplicaciones y Desarrollos de la Teoría de las Funciones Semióticas* (pp. 13 – 36). Jaén: Universidad de Jaén.

Batanero, C. y Díaz, C. (2006). Methodological and Didactical Controversies around Statistical Inference. *Actes du 36iémes Journées de la Societé Française de Statistique.* CD ROM. Paris: Societé Française de Statistique.

Batanero, C. y Díaz, C. (En prensa). Meaning and understanding of mathematics. The case of probability. En J. P Van Bendegen y K. Fraçois (Eds), *Philosophical Dimensions in Mathematics Education.* Nueva York: Springer.

Batanero, C., Díaz, C. y de la Fuente, I. (En prensa). Alcune consideración sull'insegnemento della probabilitá condizionata. *Nuova Secondaria.*

Bauersfeld, H. (1995).The structuring of the structures: Development and function of mathemathizing as a social practice. En L. Steffe y J. Gale (Eds.), *Constructivism in Education.* (pp. 137-158). Hillsdale, NJ: Lawrence Erlbaum.

Beltrán, M. (Ed.) (2001). *Actas de las Jornadas Europeas de Enseñanza y Difusión de la Estadística.* Mallorca: Instituto Balear de Estadística.

Ben-Zvi, J. y Garfield, J. (2004) (Eds.), *The challenge of developing statistical literacy, reasoning and thinking.* Dordrecht: Kluwer.

Bernard, J. M. (1998). Bayesian inference for categorised data. En H. Rouanet et al. (Eds.), *New ways in statistical methodology* (pp. 159 – 226). Berna: Peter Lang.

Bernardo, J. M. (1981). *Bioestadística. Una perspectiva bayesiana.* Barcelona: Vicens-Vives.

Bernardo, J. M. (2003). Bayesian Statistics. En R. Viertl (Ed.), *Encyclopaedia of Life Support Systems (EOLSS). Probability and Statistics*. Oxford, UK: UNESCO. On line: http://www.uv.es/~bernardo/BayesStat.pdf.

Bernardo, J. M. (2006). A Bayesian Mathematical Statistics Primer. En A. Rossman y B. Chance, (Eds.), *Proceedings.of the Seventh International Conference on Teaching Statistics.* CD ROM. Salvador de Bahia: International Association for Statistical Education.

Bernardo, J. M. y Smith, A. F. M. (1994). *Bayesian Theory*. Nueva York: Wiley.

Berry, D. A. (1995). *Basic statistics: A Bayesian perspective.* Belmont, CA: Wadsworth.

Biehler, R. (1997a). Software for learning and for doing statistics. *International Statistical Review*, *65(2),* 167-190.

Biehler, R. (1997b). Students' difficulties in practicing computer-supported data analysis: some hypothetical generalizations from results of two exploratory studies. En J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics* (pp. 176-197). Voorburg, The Netherlands: International Statistical Institute.

Biehler, R. (2001). Developing and Assessing Students' Reasoning in Comparing Statistical Distributions in Computer Supported Statistics Courses*. Presented at Statistics Literacy and Reasoning Research Forum 3. Armidale, Australia.*

Birnbaum, I. (1982). Interpreting statistical significance. *Teaching statistics. 4,* 24 – 27.

Bisquerra, R. (1989). *Métodos de investigación educativa.* Barcelona: PPU.

Black, M. (1979). *Inducción y probabilidad*. Madrid: Cátedra.
Bloom, B. S. (1956). *Taxonomy of educational objectives. Handbook 1: The cognitive domain.* New York: McGraw Hill.
Bolívar, A. (1998). Tiempo y contenido del discurso curricular en España. *Revista de currículum y formación del profesorado, 2(2).* On line: http://www.ugr.es /~recfpro /rev22ART4.pdf.
Bolstad, W. (2004). *Introduction for Bayesian statistics.* Nueva York: Wiley.
Botella, J., León, O. G. y San Martín, R. (1993). *Análisis de datos en Psicología I.* Madrid: Pirámide.
Boldstad, W. M. (2002). Teaching bayesian statistics to undergraduates: Who, what, where, when, why, and how. En B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics.* CD ROM. Ciudad del Cabo, Sudafríca: International Association for Statistics Education.
Borges, A., San Luis, C., Sánchez, J. A. y Cañadas, I. (2001). El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa. *Psicothema, 13 (1),* 174-178.
Borges, A. y Sánchez Bruno, A. (2004). Algunas consideraciones metodológicas relevantes para la investigación aplicada. *Revista Electrónica de Metodología Aplicada, 9 (1),* 1-11.
Box, G. P. y Tiao, G. C. (1992). *Bayesian inference in statistical analysis*. Nueva York: Wiley.
Brousseau, G. (1983). Les obstacles epistémologiques et les problémes en mathématiques. *Recherches en Didactique des Mathematiques, 4(2),* 164 – 198.
Brousseau, G. (1997). *Theory of didactical situations in mathematics*. Dordrecht: Kluwer.
Cabriá, S. (1994). *Filosofía de la estadística.* Valencia: Servicio de Publicaciones de la Universidad.
Canavos, G. C. (1992). *Probabilidad y Estadística. Aplicaciones y métodos*. Méjico: Mc Graw Hill.
Carmines, E. G. y Zeller, R. A. (1979). *Reliability and validity assesment*. Londres: Sage.
Carmona, J. (2004). Una revisión de las evidencias de fiabilidad y validez de los cuestionarios de actitudes y ansiedad hacia la estadística. *Statistics Education Research Journal*. *3(1).* On line. Disponible en: www.stat.auckland.ac.nz/~iase /serj/SERJ3(1) _marquez.pdf.
Castro-Posada, J. (2001). *Metodología de la investigación*. *Fundamentos.* Salamanca: Amaru.
Catena, A. Ramos, M. M. y Trujillo, H. M. (2003). *Análisis multivariado. Un manual para investigadores.* Madrid: Biblioteca nueva.
Chow, L. S. (1996). *Statistical significance: Rationale, validity and utility.* Londres: Sage.
Cohen, J. (1990). Things I have learnt so far. *American Psychologist, 45,* 1304 - 1312.
Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49(12),* 997 – 1003.
Congdon, P. (2003). *Applied Bayesian Modelling.* Nueva York: Wiley.
Cook, T. y Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field set tings.* Chicago: Rand McNally Publishing Company.
Corroyer, D. y Wolff, M. (2003). *L'analyse statistique des donneés en psychologie. Concepts et méthodes de base*. París: Armand Colin.
Cosmides, L. y Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*, 1-73.
Couturier, R. (2001). Subjects categories contribution in the implicative and the similarity analysis. En A. Gagatsis (Ed), *Learning in mathematics and science and educational technology (Vol. 2),* (pp. 369-376). Nicosia: Universidad de Chipre.
Couturier, R. y Gras, R. (2005). CHIC: Traitement de données avec l'analyse implicative. En G. Ritschard y C. Djeraba (Eds.), *Journées Extraction et Gestion des Connaissances (EGC'2005) (Vol. 2),* (pp. 679-684). Paris: Universidad de Lille.
Couturier, R., Gras, R. y Guillet, F. (2004). Reducing the number of variables using implicative analysis. En D. Banks, L. House, F. R. McMorris, P. Arabie y W. Gaul (Eds.), *Classification, Clustering, and Data Mining Application. Proceedings of the Meeting of the International Federation of Classification Societies Conference* (pp.277—285). Chicago: Springer-Verlag.
Cronbach, L. J. (1988). Five perspectives on the validity argument. En H. Wainer y H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.
Cuadrras, C. M. (1981). *Análisis multivaraitne*: Barcelona: Eudeba.
Cuadras, C. M., Echevarría B., Mateo, J. y Sánchez, P. (1984). *Fundamentos de estadística. Aplicación a las ciencias humanas.* Madrid: Promociones Publicaciones Universitarias.
Cumming, G., Williams, J. y Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299-311.
Curtis, C. (Ed.) (2002). *Actas de las Jornadas Interamericanas de Enseñanza de la Estadística.* CD ROM. Buenos Aires: Universidad Nacional Tres de Febrero..
Dane, F. C. (1990). *Research methods*. Pacific Grove, CA: Brooks/Cole.
David, J., Guillet, F., Philipp, V. y Gras, R. (2005). Implicative statistical analysis applied to clustering of terms taken from a psychological text corpus. *Applied Stochastic Models and Data Analysis (ASMDA 2005)*. On Line http://asmda 2005.enst-bretagne.fr/IMG/pdf/proceedings/201.pdf.

Davidson, R. y Swift, J. (Eds.) (1988). *Proceedings of the 2nd International Conference on Teaching Statistics.* British Columbia, Canada: International Association for Statistics Education.

De la Fuente, E. I. y Díaz, C. (2003). Reflexiones sobre los métodos inferenciales en psicología. *Libro de resúmenes del VIII Congreso de Metodología de las Ciencias Sociales y de la Salud* (pp. 326 – 327). Valencia: Departamento de Metodología de las Ciencias del Comportamiento y Asociación Española de Metodología de las Ciencias del Comportamiento.

De la Fuente, E. I., Díaz, C. y Cañadas, G. (2005). Algunas razones para introducir la inferencia bayesiana en la formación metodológica en el campo de la psicología. En I. de la Fuente et al. (Eds.), *IX Congreso de Metodología de las Ciencias Sociales y de la Salud. Libro de resúmenes* (p. 104). Granada: Universidad de Granada.

De la Fuente, E. I., García, J. y De la Fuente, L. (2002). Estadística Bayesiana en la Investigación Psicológica. *Metodología de las Ciencias del Comportamiento*, *4*, 185-200.

De Groot, M. H. (1988). *Probabilidad y estadística*. Delaware: Addisson Wesley.

Delgado, J. M. y Gutiérrez, J. (1994) *Métodos y técnicas cualitativas de investigación en ciencias sociales*. Madrid: Síntesis Psicología.

DelMas, R. C., Garfield, J. B. y Chance, B. L. (1998). Exploring the role of computer simulations in developing understand of sampling distributions. Trabajo presentado en el *American Educational Research Association*. *Annual Meeting*. Montreal.

Díaz, C. (2003). Heurísticas y sesgos en el razonamiento probabilístico. Implicaciones para la enseñanza de la estadística. *Actas del 27 Congreso Nacional de Estadística e Investigación Operativa*. CD ROM. Lleida: Universidad de Lleida

Díaz, C. (2004). *Elaboración de un instrumento de evaluación del razonamiento condicional. Un estudio preliminar*. Trabajo para la obtención del Diploma de Estudios Avanzados. Universidad de Granada.

Díaz, C. (2005). Evaluación de la falacia de la conjunción en alumnos universitarios. *Suma, 48,* 45-50.

Díaz, C. y Batanero, C. (2005). La probabilidad condicional en los textos de estadística para psicología. *Actas del V CIBEM, Congreso Iberoamericano de Educación Matemática.* CD ROM. Oporto: Sociedad Portuguesa de Profesores de Matemáticas.

Díaz, C., Batanero, C. y Cobo, B. (2003). Fiabilidad y generalizabilidad. Aplicaciones en evaluación educativa. *Números, 54*, 3 – 21.

Díaz, C. y de la Fuente, E. I. (2004). Controversias en el uso de la inferencia en la investigación experimental. *Metodología de las Ciencias del Comportamiento, (Volumen especial 2004)*, 161-167.

Díaz, C. y de la Fuente, E. I. (2005a). Recursos para la enseñaza del razonamiento condicional en Internet. Trabajo presentado en el *Congreso Internacional* "*El Profesorado ante el reto de las Nuevas Tecnologías en la Sociedad del Conocimiento"*. Universidad de Granada.

Díaz, C., y de la Fuente, I. (2005b). Razonamiento sobre probabilidad condicional e implicaciones para la enseñanza de la estadística. *Epsilon, 59*, 245-260.

Díaz, C., y de la Fuente, E. I. (2005c). Construcción de un cuestionario sobre comprensión de la probabilidad condicional. En J. Ortiz y A. Montenegro (Eds.), *Actas del XV Simposio de Estadística.* CD ROM. Bogotá: Universidad Nacional de Colombia.

Díaz, C. y de la Fuente, E. I. (2005c). Conflictes semiòtics en el càlcul de probabilitats a partir de taules de doble entrada. *Biaix*, *24,* 85-91.

Díaz, C. y de la Fuente, E. I. (2006). Assessing psychology students' difficulties with conditional probability and bayesian reasoning. En A. Rossman y B. Chance (Eds.), *Proceedings of ICOTS – 7.* CD ROM. Salvador de Bahia: International Association for Statistical Education.

Díaz, C. y de la Fuente, E. I. (En prensa). Dificultades en la resolución de problemas bayesianos: un estudio exploratorio en estudiantes de psicología. *Educación Matemática.*

Díaz, C., de la Fuente, E. I. y Batanero, C. (2004a). Statistical inference and experimental research. Should we revise our educational practices? *Libro de resúmenes de ICME-10.* Copenhague, Dinamarca: ICMI.

Díaz, C., de la Fuente, E. I. y Batanero, C. (2004b). Competencia de estudiantes de psicología en la resolución de problemas bayesianos. *Libro de resúmenes de la XVIII Reunión Latinoamericana de Matemáticas Educativa*. Tuxtla Gutiérrez, México: Comité Latinoamericano de Matemática Educativa.

Díaz, C., de la Fuente, E. I. y Wihelmi, M. (En prensa). Implications between learning outcomes in elementary bayesian inference. En R. Gras (Ed.), *Statistical Implicative Analysis: theory and applications*. Nueva York: Springer.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. En D. Kahneman, P. Slovic y Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases.* Nueva York: Cambridge University Press.

Edwards, W., Lindman, H. y Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242.

Einhorn, H. J. y Hogart, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3 – 19.

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

Engel, J. (2002). Activity-based statistics, computer simulation and formal mathematics. En Phillips, B. (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics.* CD ROM. Ciudad del Cabo, Sudáfrica: IASE.

Engel, J. (2003). *Statistics education and the Internet.* Berlin: International Association for Statistical Education.

Engel, J. y Vogel, M. (2004). Mathematical Problem Solving as Modelling Process. En H. Henn y W. Blum (Eds.), *ICMI-Study 14: Application and Modeling in Mathematics Education*. Dormunt, Alemania: ICMI.

Ernest, P. (1994). Varieties of constructivism: Their metaphors, epistemologies and pedagogical implications. *Hiroshima Journal of Mathematics Education 2,* 1-14.

Ernest, P. (1998). *Social constructivism as a philosophy of mathematics*. New York: SUNY.

Estepa, A. (1993). *Concepciones iniciales sobre la asociación estadística y su evolución como consecuencia de una enseñanza basada en el uso de ordenadores*. Tesis doctoral. Universidad de Granada.

Estes, W. K. (1997). Significance testing in psychological research: Some persisting issues. *Psychological Science*, *8(1),* 18 – 20.

Estrada, A. y Díaz, C. (2006). Computing probabilities from two way tables. An exploratory study with future teachers. En A. Rossman y B. Chance (Eds.), *Proceedings of ICOTS-7*. CD ROM. Salvador (Bahia): International Association for Statistical Education.

Estrada, A., Díaz, C. y de la Fuente, E. I. (2006). Un estudio inicial de sesgos en el razonamiento sobre probabilidad condicional en alumnos universitarios. *Actas del IX Simposio de la SEIEM.* Huesca: Sociedad Española de Investigación en Educación Matemática.

Falk, R. (1986). Conditional probabilities: insights and difficulties. En R. Davidson y J. Swift (Eds.), *Proceedings of the Second International Conference on Teaching Statistics* (pp. 292 – 297). British Columbia, Canadá: University of Victoria.

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

Falk, R. (1989). Inference under uncertainty via conditional probability. En R. Morris (Ed.), *Studies in mathematics education,* vol. 7 (pp. 175 – 184). Paris: UNESCO.

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

Feller, W. (1973). *Introducción a la teoría de las probabilidades y sus aplicaciones.* Méjico: Limusa.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research, 50,* 123-129.

Fidler, F. (2002). The fifth edition of the APA publication manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, *62 (5),* 749-770.

Finch, S., Cumming, G., y Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181-210.

Fisher, R. A. (1956). Mathematics of a lady testing tea, En J. Newman (Ed.), *The world of mathematics Vol., III.* Simon and Schuster (1979). Traducido como Las matemáticas de la catadora de té. En J. R. Newman (Ed.), *El mundo de las matemáticas Vol. 3*, (pp. 194 – 203). Barcelona: Grijalbo,

Frías, M. D., Pascual, J. y García, J. F. (2000). Tamaño del efecto del tratamiento y significación estadística. *Psicotema, 12 (2)*, 236-240.

Frías, M. D., Pascual, J. y García, J. F. (2002). La hipótesis nula y la significación práctica. *Metodología de las Ciencias del Comportamiento, 4 (especial),* 181-185.

Fox, D. J. (1981). *El proceso de investigación en la educación*. Pamplona: Eunsa.

Gal, I. y Garfield, J. (Eds.) (1997). *The assessment challenge in statistics education*. The Netherland: IOS Press.

Galmacci, G. (2001). The impact of Internet on the researchers' training. En C. Batanero (Ed.), *Training researchers in the use of statistics* (pp. 159-169). Granada: International Statistical Institute.

Garfield, J. B. y Burrill, G. (Eds.) (1997). *Research on the role of technology in teaching and learning statistics.* Voorburg: International Association for Statistical Education e International Statistical Institute.

Gelman, A., Carlin, J. B. Stern, H. S. y Rubin, D. B. (2003). *Bayesian data analysis*. Londres: Chapman and Hall.

Ghiglione, R. y Matalón, B. (1991). *Les enquêtes sociologiques. Théories et practique.* París: Armand Colin.

Gil Flores, J. (1994). *Análisis de datos cualitativos*. Barcelona. P.P.U.

Gingerenzer, G. (1993). The superego, the ego and the id in statistical reasoning. En G. Keren y C. Lewis (Eds.), *A handbook for data analysis in the behavioural sciences: Methodological issues* (pp. 311 – 339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice-versa). En G. Wright y P. Ayton (Eds.), *Subjective probability* (pp. 129 – 161). Chichester: Wiley.

Gigerenzer, G. y Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review,* 102, 684 – 704.

Gingerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. y Kruger, L. (1989). *The empire of chance. How probability changed science and everyday life*. Cambridge: Cambridge University Press.

García Cueto, E. (1993). *Introducción a la psicometría*. Madrid: Siglo XXI.

Glass, G. V. y Stanley, J. C. (1974). *Métodos estadísticos aplicados a las ciencias sociales.* Méjico: Prentice Hall.

Goetz, J. P. y Lecompte, M. D. (1988). *Etnografía y diseño cualitativo en investigación educativa*. Madrid: Morata.

Granaas, M. (2002). Hypothesis testing in psychology: throwing the baby out with the bathwater? En B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics.* CD ROM. Ciudad del Cabo, Sudáfrica: IASE.

Gras, R. y Totohasina, A. (1995). Chronologie et causalité, conceptions sources d'obstacles épistémologiques à la notion de probabilité conditionnelle. *Recherches en Didactique des Mathématiques*. *15(1),* 49 – 95.

Gras, R. (1993). Une méthode de classification non symétrique: l'implication statistique. *Boulletin de la Société Française de Classification*, 1.

Gras, R. (1996). *L'implication stastistique : nouvelle méthode exploratoire de donnés applications a la didactique*. Grenoble: La Pensée Sauvage.

Gras, R., Diday, E., Kuntz, P. y Couturier, R. (2001). Variables sur intervalles et variables intervalles en analyse statistique implicative. En *Actes de Société Francophone de Classification, SFC'2001*, (pp 166—173). Guadeloupe, France: Société Francophone de Classification.

Gras, R., Kuntz P. y Briand, H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et Sciences Humaines*, *154-155*, 9-29.

Gras, R. y Ratsima-Rajohn, H. (1996). L'implication statistique, une nouvelle méthode d'analyse de données. *Recherche Opérationnelle, 30 (3),* 217-232.

Grey, D. R. (Ed.) (1982). *Proceedings of the First International Conference on Teaching Statistics.* Sheffield: Centre for Statistical Education, University of Sheffield.

Godino, J. D. (2005). *Marcos teóricos de referencia sobre la cognición matemática*. Documento de trabajo del curso de doctorado "Teoría de la educación Matemática". On line: http://www.ugr.es/local/jgodino.

Hager, W. (2000). About some misconceptions and the discontent with statistical tests in psychology. *Methods on Psychological Research, 5(1).* On line: http://www.mpr-online.de.

Hagod, M. J. (1970). The notion of hypothetical universe. En D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 65 – 79). Chicago: Aldine.

Hamerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology, 101,* 252 – 254.

Harlow, L. L. (1997). Significance testing: Introduction and overview. En L. L. Harlow, S. A. Mulaik, y J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 1 – 20). Mahwah, NJ: Erlbaum.

Harlow, L. L., Mulaik, S. A. y Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science. 8(1),* 8 – 11.

Hawkins, A. (1997). How far have we come? Do we know where we are going? En E. M. Tiit (Ed.), *Computational statistics & statistical education* (pp. 100-122). Tartu, Estonia: International Association for Statistical Education e International Association for Statistical Computing.

Hawkins, A. (1999). What is the International Statistical Institute? *Teaching Statistics, 21(2),* 34 – 35.

Heitele, D. (1975). An epistemological view on fundamental stochastic ideas. *Educational Studies in Mathematics, 6,* 187 – 205.

Hernández, R., Fernández, C. y Baptista, C. (1998). *Metodología de investigación.* Méjico: McGraw-Hill.

Hertwig, R. y Gigerenzer, G. (1999). The conjunction fallacy revisited: how intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making, 12(4),* 275.

Hiebert, J. y Lefevre, P. (1987). Conceptual and procedural knowledge in mathematics: An introductory analysis. En J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics*. London: LEA Publishers.

Holmes, P. (2002). Some lessons to be learnt from curriculum developments in statistics. En B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics.* CD ROM. Ciudad del Cabo, Sudáfrica: IASE.

Huberman, A. M. y Miles, M. (1994). Data management and analysis methods. En N. K. Denzin y Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 428 – 444). London: Sage Publications.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8 (1),* 3 – 7.

Iglesias, P., Leiter, J., Mendoza, M., Salinas, V. y Varela, H. (2000). Mesa redonda sobre enseñanza de la estadística bayesiana. *Revista de la Sociedad Chilena de Estadística, 16-17*, 105-120.

Ito, P. K. (1999). Reaction to invited papers on statistical education and the significance tests controversy. *Proceedings of the Fifty-second Session of the International Statistical Institute (Tome 58, Book 3)* (pp. 101-103). Helsinki, Finlandia: International Statistical Institute.

Iversen, G. R. (1998), Student Perceptions of Bayesian Statistics. En Pereira- Mendoza (Ed.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 234-240). Singapore: International Statistical Institute.

Jones, G. A. (2005) (Ed.), *Exploring probability in school: Challenges for teaching and learnin.* Nueva York: Springer.

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

Kaput, J. (1991). Notations and representations as mediators of constructive processes. En E. von Glasersfeld (Ed.), *Constructivism and mathematics education* (pp. 53–74). Boston: Reidel.

Kelly, I. W. y Zwiers, F. W. (1986). Mutually exclusive and independence: Unravelling basic misconceptions in probability theory. *Teaching Statistics, 8,* 96 – 100.

Kilpatrick, J. (1992). A history of research in mathematics education. En D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 3-38). Nueva York: MacMillan.

Kirk, J y Miller, M. L. (1986). *Reliability and validity in qualitative research.* Londres: Sage.

Kish, L. (1970). Some statistical problems in research design. En D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 127 – 141). Chicago: Aldine.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavior and Brain Sciences*, *19,* 1-54.

Kurzenhauser, S. y Hoffrage, U. (2002). Teaching Bayesian reasoning: An evaluation of a classroom tutorial for medical students. *Medical Teacher*, *24 (5),* 516 – 521.

Labovitz, S. (1970). Criteria for selecting a significance level: A note on the sacredness of .05. En D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 166-170). Chicago: Aldine.

Lagrange J. B. (1998). Analyse implicative d'un ensemble de variables numériques. *Revue de Statistique Appliquée, 46 (1),* 71-93.

Lahanier-Reuter, D. (2001). Grouping together variables values: an algorithm in implicative analysis. *Mathematique et Sciences Humaines*, *154,* 47-59.

Lawrence, J. (2003). *A quick introduction to First Bayes.* Montreal: Mc Gill University. On line*:* http://www.medicine.mcgill.ca/epidemiology/Joseph/pdf/First.Bayes.pdf.

Lecoutre, B. (1996). *Traitement statistique des données experimentales.* París: CISIA.

Lecoutre, B. (1999). Beyond the significance test controversy: Prime time for Bayes? *Bulletin of the International Statistical Institute: Proceedings of the Fifty-second Session of the International Statistical Institute (Tome 58, Book 2)* (pp. 205 – 208). Helsinki, Finlandia: International Statistical Institute.

Lecoutre, B. (2000). From signicance tests to .fiducial Bayesian inference. En H. Rouanet, J.M. Bernard, M.C. Bert, B. Lecoutre, M.P. Lecoutre y B. Le Roux (Eds.), *New ways in statistical methodology. From significance tests to Bayesian inference. (2nd edition),* (pp. 123-157). Paris: Peter Lang.

Lecoutre, B. (2006). Training students and researchers in Bayesian methods for experimental data analysis. *Journal of Data Science*, *4(2),* 207-232.

Lecoutre, B., Lecoutre M.P., y Poitevineau J. (2001). Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *International Statistical Review*, *69,* 399-418.

Lee, P. M. (2004). *Bayesian statistics. An introduction.* York, UK : Arnold.

León, O. G. y Montero, I. (2002). *Métodos de investigación en psicología y educación.* Madrid: McGraw-Hill.

Leonard, T. y Hsu, J. S. (2001). *Bayesian methods*. Cambridge: Ganbrodge Universisty Press.

Lerman, I. C. (1981). *Classification et analyse ordinale des données*. Paris: Dunod.

Lerman, I. C; Gras, R. y Rostam, H. (1981a). Elaboration d'un indice d'implication pour données binaires I. *Mathématiques et sciences humaines, 74,* 5-35.

Lerman, I. C; Gras, R. y Rostam, R. (1981b). Elaboration d'un indice d'implication pour données binaires II. *Mathématiques et sciences humaines, 75,* 5-47.

Levin, J. R. (1998a). To test or not to test H_0? *Educational and Psychological Measurement, 58,* 313 – 333.

Levin, J. R. (1998b). What if there were no more bickering about statistical significance tests? *Research in the Schools*, *2,* 45 – 53.

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

Lindley, D. L. (1993). The analysis of experimental data. The appreciation of tea and wine. *Teaching Statistics, 15 (1),* 22-25.

Lipset, S. M., Trow, M. A. y Coleman, J. S. (1970). Statistical problems. En D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 81 – 86). Chicago: Aldine.

Lonjedo, M. A. y Huerta, P. (2004). Una clasificación de los problemas escolares de probabilidad condicional. Su uso para la investigación y el análisis de textos. En Castro, E., y De la Torre, E. (Eds.), *Investigación en Educación Matemática. Octavo Simposio de la Sociedad Española de Investigación en Educación Matemática*, (pp 229-238). A Coruña: Universidade da Coruña.

Lonjedo, M, A. y Huerta, P. (2005). The nature of the quantities in a conditional probability problem. Its influence in the problem resolution. *Proceedings of CERME IV*. On line: http://cerme4.crm.es/Papers% 20definitius/5/wg5litofpapers.

Losada, J. L. y López- Feal, R. (2003). *Métodos de investigación en ciencias humanas y sociales.* Madrid: Thompson.

López Feal, R. (1986). *Construcción de instrumentos de medida en ciencias conductuales y sociales.* Barcelona: Alamex.

Martínez Arias, R. (1995). *Psicometría: teoría de los tests psicológicos y educativos.* Madrid: Síntesis.

Martínez Bonafé, J. (1995). Interrogando al material curricular (Guión para el análisis y la elaboración de materiales para el desarrollo del curriculum). En J. García y M. Beas (Eds.), *Libro de Texto y construcción de materiales curriculares* (pp.221 – 240). Granada: Proyecto Sur de Ediciones.

Martignon, L. y Wassner, C. (2002). Teaching decision making and statistical thinking with natural frequencies. En B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics.* CD ROM. Ciudad del Cabo, Sudáfrica: IASE.

Matthews, R. A. (1998). Facts versus factions: The use and abuse of subjectivity in scientific research. European Science and Environment Forum Working Paper. En J. Morris (Ed.), *Rethinking risk and the precautionary principle* (pp. 247-282). Oxford: Butterworth.

Maury, S. (1985). Influence de la question dans une épreuve relative á la notion d'independance. *Educational Studies in Mathematics, 16,* 283 – 301.

Maury, S. (1986). *Contribution a l'étude didactique de quelques notions de probabilité et de combinatoire à travers la résolution de problémes.* Tesis doctoral. Universidad de Montpéllier II.

McLean, A. (2001). Statistics in the catwalk. The importance of models in training researchers in statistics. En C. Batanero (Ed), *Training Researchers in the Use of Statistics.* Granada: International Association for Statistics Education and International Statistical Institute.

Meliá, J. L. (2001). *Teoría de la fiabilidad y la validez*. Valencia: Cristóbal Serrano.

Menon, R. (1993). Statistical significance testing should be discontinued in mathematics education research. *Mathematics Education Research Journal, 5(1),* 4 – 18.

Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational Measurement*. 3ª *Ed*. (pp. 13-103). Nueva York: Collier Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into scoring meaning. *American Psychologist, 9*, 741-749.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45,* 35-44.

Millman, J. y Greene, J. (1989). The specification and development of test of achivement and ability. En R. L. Linn (Ed.), *Educational Measurement* (pp. 335 – 366). Londres: Macmillan.

Mises, R. von (1952). *Probability, statistics and truth.* J. Neyman, O. Scholl, y E. Rabinovitch, (Trans.). Londres: William Hodge &co. (Original work published 1928).

Molinero, A. (2002). *El método bayesiano en la investigación médica.* Madrid: Asociación española contra la hipertensión arterial. On line: http://www.seh-lelha.org/ bayes1.htm.

Monterde-Bort, H., Pascual, J. y Frías, M. D. (2005). Incomprensión de los conceptos metodológicos y estadísticos: La encuesta USABE. Trabajo presentado en el *IX Congreso de Metodología de las Ciencias Sociales y de la Salud.* Granada.

Monterde-Bort, H., Pascual, J. y Frías, M. D. (En prensa). Errores de interpretación de los métodos estadísticos: importancia y recomendaciones. *Psicothema.*

Moore, D. S. (1992). Teaching Statistics as a respectable subject. En F. Gordon y S. Gordon (Eds.), *Statistics for the Twenty-First Century* (pp. 14-25). The Mathematical Association of America.

Moore, D. S. (1995). *Estadística aplicada básica*. Barcelona: Antoni Bosch.

Moore, D. S. (1997a). New pedagogy and new content: The case of statistics. *International Statistical Review, 65(2),* 123 – 155.

Moore, D. S. (1997b). Bayes for beginners? Some pedagogical questions. En S. Panchapakesan (Ed.), *Advances in Statistical Decision Theory* (pp. 3-17). Birkhäuser.

Moore, D. S. (1997c). Bayes for beginners? Some reason to hesitate. *The American Statistician*, *51(3),* 254-261.

Morales, P. (1988). *Medición de actitudes en psicología y educación*. San Sebastián: Universidad de Comillas.

Moses, L. E. (1992). The reasoning of statistical inference. En D. C. Hoaglin y D. S. Moore (Eds.), *Perspectives on contemporary statistics* (pp. 107 – 122). Washington, DC: Mathematical Association of America.

Morrison, D. E., y Henkel, R. E. (Eds.) (1970). *The significance tests controversy. A reader*. Chicago: Aldine.
Muñiz, J. (1994). *Teoría clásica de los tests.* Madrid: Pirámide.
Murphy, K. R. y Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimun-effect tests in the general linear model. *Journal of Applied Psychology*, *84*, 234-2484.
Nisbett, R., y Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgments*. Englewood Cliffs, NJ: Prentice Hall.
Nortes Checa, A. (1993). *Estadística teórica y aplicada*. Barcelona: PPU.
National Organising Committee ICOTS-4 (1994). *Proceedings of the 4th International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute.
Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, UK: Wiley.
O'Hagan, A. y Forster, J. (2004). *Kendall's Advanced Theory of Statistics. Bayesian Inference. Vol. 2B*. Londres: Arnold.
Ojeda, A. M. (1995). Dificultades del alumnado respecto a la probabilidad condicional. *UNO, 5,* 37 – 55.
Ortega, A. R. (1991). *Contingencia y juicios de covariación en humanos.* Granada: Servicio de Publicaciones de la Universidad de Granada.
Ortega, A. R. (1999). Aproximación histórica al análisis de datos en Psicología desde la estadística. En M. Román (Ed.), *Educar enseñando: antología de estudios científicos en homenaje a la profesora Mercedes Lamarque Forn*, (pp. 253-272). Jaén: Universidad de Jaén.
Ortega, A. I., Martos, R. y Ortega, A. R. (1992). Juicios de contingencia. *Revista de la Facultad de Humanidades de Jaén, 1 (3),* 115-140.
Osterlind, S. J. (1989). *Constructing test items.* Boston: Kluwer.
Pascual, J., García, J. F. y Frías, M. D. (2000). Significación estadística, importancia del efecto y replicabilidad de los datos. *Psicothema, 12 (2)*, 408-412.
Pedhazur, E. J. y Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum.
Peña, D. (1986). *Estadística. Modelos y Métodos 1. Fundamentos.* Madrid: Alianza Editorial.
Peña, D. (2002). *Análisis de datos multivariables.* Madrid: McGraw-Hill.
Peña. D. y Romo, J. (1997). *Introducción a la estadística para las ciencias sociales.* Madrid: McGraw-Hill.
Pereira-Mendoza, L., Seu Kea, L., Wee Kee, T. y Wong, W.K. (Eds.) (1998). *Statistical Education-Expanding the Network. Proceedings of the Fifth International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute.
Pérez Echeverría, M. P. (1990). *Psicología del razonamiento probabilístico.* Madrid: Ediciones de la Universidad Autónoma de Madrid.
Phillips, B. (Ed.) (1996). *Papers on Statistical Education presented at ICME-8.* Swinburne, Australia: University of Technology.
Phillips, B. (Ed.) (2002). *Proceedings of the Sixth International Conference on Teaching of Statistics*. CD ROM. Ciudad del Cabo, Sudáfrica: IASE.
Piaget, J. (1979). *Introducción a la epistemología genética*. Buenos Aires: Paidos.
Piaget, J. e Inhelder, B. (1951). *La genèse de l'idèe de hasard chez l'enfant.* Paris: Presses Universitaires de France.
Poitevineau, J. (1998). *Méthodologie de l'analyse des données expérimentales: étude de la pratique des tests statistiques chez les chercheurs en psychologies: approches normative, prescriptive et descriptive*. Tesis doctoral. Universidad de Rouen.
Pollard, P., y Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 10,* 159 – 163.
Pollatsek, A., Well, A. D., Konold, C. y Hardiman, P. (1987). Understanding conditional probabilities. *Organitation, Behavior and Human Decision Processes, 40,* 255 – 269.
Popper, K. R. (1967). *La lógica de la investigación científica.* Madrid: Tecnos.
Pruzek. R. M. (1997). An introduction to bayesian inference and its applications. En L. L. Harlow, S. A. Mulaik y J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 287-318). Mahwah, NJ: Lawrence Erlbaum.
Ramos, M., Catena, A. y Trujillo, H. (2004). *Manual de métodos y técnicas de investigación en ciencias del comportamiento*. Madrid: Biblioteca Nueva.
Rios, S. (1967). *Métodos estadísticos.* Madrid: Ediciones del Castillo.
Rindskopf, D. M. (1997). Classical and bayesian approaches. En L. L. Harlow, S. A. Mulaik y J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 319-334). Mahwah, NJ: Lawrence Erlbaum Associates.
Ritschard, G. (2005). De l'usage de la statistique implicative dans les arbres de classification, Trabajo presentado en las *Jounrnées ASI'05.* Palermo.
Rivadulla, A. (1991). *Probabilidad e inferencia científica.* Barcelona: Anthropos.

Robert, C. P. (2001). *The bayesian choice*. Nueva York: Springer.
Rossman, A. y Chance, B. (2006). *Proceedings of the Seventh International Conference on Teaching Statistics.* CD ROM. Salvador de Bahia, Brasil: International Association for Statistical Education.
Rossman, A., & Short, T. (1995). Conditional probability and education reform: Are they compatible? *Journal of Statistics Education, 3 (2).* On line: http://www.amstat. org/publications/jse/v3n2/rossman.html.
Rouanet, H. (1998a). Statistics for researchers. En H. Rouanet et al. (Eds.), *New ways in statistical methodology* (pp. 1 – 28). Berna: Peter Lang.
Rouanet, H. (1998b). Statistical practice revisited. En H. Rouanet et al. (Eds.), *New ways in statistical methodology* (pp. 29 – 64). Berna: Peter Lang.
Rozeboon, W. W. (1970). The fallacy of the null hypothesis significance test. En. D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 216 – 230). Chicago: Aldine.
Sacerdote, A. y Balima, G. (En preparación). *Estadística bayesiana.* Buenos Aires: Universidad de Buenos Aires. On line: http://www.fi.uba.ar/materias/6109 /libro.html.
Sánchez, E. (1996). Dificultades en la comprensión del concepto de eventos independientes. En F. Hitt (Ed.), *Investigaciones en Educación Matemática* (pp. 389 – 404). México: Grupo Editorial Iberoamérica.
Sánchez, E. y Hernández, R. (2003). Variables de tarea en problemas asociados a la regla del producto en probabilidad. En E. Filloy (Coord.), *Matemática educativa, aspectos de la investigación actual* (pp. 295 – 313). México: Fondo de Cultura Económica.
Santiesteban, C. (1990). *Psicometría aplicada.* Madrid: Norma.
Sax, G. (1989). *Principles of educational and psychological measurement and* evaluation. Belmont, CA: Wadsworth.
Scholz, R. W. (1991). Psychological research in probabilistic understanding. En R. Kapadia, & M. Borovcnik (Eds.)*, Chance encounters: Probability in education* (pp. 213 – 249). Dordrecht: Kluwer.
Schuyten, G. (1991). Statistical thinking in psychology and education. En D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (pp. 486 – 490). Voorburg, The Netherlands: International Statistical Institute.
Sedlmeier, P. (1999). *Improving statistical reasoning. Theoretical models and practical implications.* Mahwah, NJ: Erlbaum.
Selander, S. (1990). Análisis del texto pedagógico. En J. García y M. Beas (Comp.), *Libro de texto y construcción de materiales curriculares,* (pp. 131 – 161). Granada: Proyecto Sur de Ediciones.
Selvin, H. C. (1970). A critique of tests of significance in survey research. En. D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 94 – 106). Chicago: Aldine.
Serrano Angulo, J. (2003). *Iniciación a la estadística bayesiana.* Madrid: La Muralla.
Sfard, A. (2000). Symbolizing mathematical reality into being -or how mathematical discourse and mathematical objects create each other. En, P. Coob, E. Yackel y K. McClain (Eds.), *Symbolizin and communicating in mathematics classrooms* (pp.38-75). Londres: Lawrence Erlbaum.
Shaughnessy, J. M. (1977). Misconceptions of probability: An experiment with a small – group, activity – based, model building approach to introductory probability at the college level. *Educational Studies in Mathematics*, *8,* 295 – 316.
Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. En D. A. Grows (Eds.), *Handbook of research on mathematics teaching and learning* (pp. 465 – 494). Nueva York: MacMillan.
Shaughnessy, J. M. (2006). Research on students' understanding of some big concepts in statistics. En G. Burrill (Ed.), *NCTM 2006 Yearbook: Thinking and reasoning with data and chance* (pp. 77-95). Reston, VA: NCTM.
Shaughnessy, J. M. (En prensa). Research on statistics learning and reasoning. En F. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning.* Greenwich, CT: Information Age Publishing, Inc., and NCTM.
Shaughnessy, J. M., Garfield, J., y Greer, B. (1996). Data handling. En A. Bishop et al. (Eds.), *International handbook of mathematics education. Vol.1* (pp. 205-237). Dordrecht, Netherlands: Kluwer.
Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory & Psychology*, *8(3),* 291-311.
Skipper, J. K., Guenter, A. L. y Nass, G. (1970). The sacredness of .05: A note concerning the uses of statistical levels of significance in social sciences. En D. E. Morrison y R. E. Henkel, (Eds.), *The significance tests controversy: A reader* (pp. 155 – 160). Chicago: Aldine.
Spatz, C. (1993). *Basic statistics. Tales of distributions.* Pacific Grove, CA: Brooks /Cole.
Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology 15*, 201-293. Disponible en Internet: http://psychclassics.yorku.ca/Spearman.
Spiegel, M. R. (1991). *Estadística.* Mc Graw Hill (2ª edición).

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter? *Theory & Psychology*, *8 (3),* 291-311.

Stangl D. K. (1998). Classical and Bayesian Paradigms: Can We Teach Both? En *Proceedings of the Fifth International Conference on Teaching Statistics.* Voorburg, The Netherlands: International Statistical Institute.

Starkings, S. (2000). *Stochastics papers presented at ICME-9.* Tokio: International Association for Statistics Education.

Stolarz-Fantino, S., Fantino, E., Zizzo, D.J. y Wen, J. (2003). The conjunction effect: New evidence for robustness. *The American Journal of Psychology, 116 (1),* 15 – 34.

Tabachnick, B. G. y Fidell, E. S. (2001). *Using multivariate statistics.* Needham Heights, MA: Allyn & Bacon.

Tarr, J. E. y Jones, G. A. (1997). A framework for assessing middle school students' thinking in conditional probability and independence. *Mathematics Education Research Journal, 9*, 39-59.

Tarr, J. E. y Lannin, J. K. (2005). How can teachers build notions of conditional probability and independence? En G. Jones (Ed,), *Exploring probability in school. Challenges for teaching and learning.* Nueva York: Springer.

Taylor, S. J. y Bogdan, R. (1986). *Introducción a los métodos cualitativos de investigación.* Buenos Aires: Paidós.

Teigen, K. H., Brun, W. y Frydenlund, R. (1999). Judgments of risk and probability: the role of frequentistic information. *Journal of Behavioral Decision Making*, *12(2),* 123.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25(2),* 26 – 30.

Thorndike, R. L. (1989). *Psicometría aplicada.* Méjico: Limusa.

Totohasina, A. (1992). *Méthode implicative en analyse de données et application á l'analyse de concetpions d'étudiants sur la notion de probabilité conditionnelle.* Tesis Doctoral. Universidad Rennes I.

Truran, J. M. y Truran, K. M. (1997). Statistical independence: One concept or two? En B. Phillips (Ed.), P*apers from Statistical Education Presented at ICME 8* (pp. 87 – 100). Swinburne: University of Technology.

Tversky, A. y Kahneman, D. (1982a). Causal schemas in judgment under uncertainty. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 117 – 128). Cambridge, MA: Cambridge University Press.

Tversky, A. y Kahneman, D. (1982b). On the psychology of prediction. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 69 – 83). Cambridge, MA: Cambridge University Press.

Tversky, A. y Kahneman, D. (1982c). Evidential impact of base rates. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 153 – 160). Cambridge, MA: Cambridge University Press.

Tversky, A. y Kahneman, D. (1982d). Judgement of and by representativeness. En D. Kahneman, P. Slovic y A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 84 – 98). Cambridge, MA: Cambridge University Press.

Vacha-Haase, T. (2001). Statistical significance should not be considered one of life's guarantees: Effect sizes are needed. *Educational and Psychological Measurement*, *61,* 219-224.

Valera, A. y Sánchez, J. (1997). Pruebas de significación y magnitud del efecto: reflexiones y propuestas. *Anales de Psicología*, *13*, 85-90.

Valera, S., Sánchez, J. y Marín, F. (2000). Contraste de hipótesis e investigación psicológica española: Análisis y propuestas. *Psicothema, 12(2),* 549-582.

Valera, A., Sánchez, J., Marín, F. y Velandrino, A.P. (1998). Potencia Estadística de la Revista de Psicología General y Aplicada (1990-1992). *Revista de Psicología General y Aplicada. 51 (2).*

Vallecillos, A (1994). *Estudio teórico experimental de errores y concepciones sobre el contraste de hipótesis en estudiantes universitarios.* Tesis doctoral. Universidad de Granada.

Vallecillos, A. (1996). Comprensión de la lógica del contraste de hipótesis en estudiantes universitarios. *Recherches en Didactique des Mathematiques, 15(3),* 53 – 81.

Vallecillos, A. (1999). Some empirical evidence on learning difficulties about testing hypotheses. *Bulletin of the International Statistical Institute: Proceedings of the Fifty-second Session of the International Statistical Institute* (Tome 58, Book 2) (pp. 201 – 204). Helsinki, Finlandia: International Statistical Institute.

Vallecillos, A. y Batanero, C. (1996). Conditional probability and the level of significance in tests of hypotheses. En L. Puig y A. Gutiérrez (Eds.), *Proceedings of the Twentieth Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 271 – 378). Valencia: Universidad de Valencia.

Van Dalen, D. B. y Meyer, W. J. (1984). *Manual de técnica de la investigación educacional*. Barcelona: Paidós.

Vere-Jones, D. (1997). The coming age of statistical education. *International Statistical Review*, *63(1),* 3 – 23.

Visauta, R. (1989). *Técnicas de investigación social I. Recogida de datos*. Barcelona: P.P.U.

Weber, R. P. (1985). *Basic content analysis.* Londres: Sage.

Western, B. (1999). Bayesian analysis for sociologists: An introduction. *Sociological Methods & Research*, *28 (1),* 7-34.

White, A. L. (1980). Avoiding errors in educational research. En R. J. Shumway (Ed.), *Research in mathematics education* (pp. 47 – 65). Reston, VA: National Council of Teachers of Mathematics.

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

Zhu, M. y Lu, A.Y. (2004). The counter-intuitive non-informative prior for the Bernoulli family. *Journal of Statistics Education, 12 (2),* On line: http://www.amstat .org/publications/jse/v12n2/zhu.pdf.