INTERNATIONAL JOURNAL OF PSYCHOLOGY, 2003, *38* (0), 000–000

Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests

Marie-Paule Lecoutre

ERIS, Université de Rouen, Mont-Saint-Aignan, France

Jacques Poitevineau

ERIS, CNRS, Paris, France

Bruno Lecoutre

ERIS, CNRS and Université de Rouen, Mont-Saint-Aignan, France

We investigated the way experienced users interpret Null Hypothesis Significance Testing (NHST) outcomes. An empirical study was designed to compare the reactions of two populations of NHST users, psychological researchers and professional applied statisticians, when faced with contradictory situations. The subjects were presented with the results of an experiment designed to test the efficacy of a drug by comparing two groups (treatment/placebo). Four situations were constructed by combining the outcome of the *t* test (significant vs. nonsignificant) and the observed difference between the two means *D* (large vs. small). Two of these situations appeared as conflicting (*t* significant/*D* small and *t* nonsignificant/*D* large). Three fundamental aspects of statistical inference of statistical inference were investigated by means of open questions: drawing inductive conclusions about the magnitude of the true difference from the data in hand, making predictions for future data, and making decisions about stopping the experiment. The subjects were 25 statisticians from pharmaceutical companies in France, subjects well versed in statistics, and 20 psychological researchers from various laboratories in France, all with experience in processing and analyzing experimental data. On the whole, statisticians and psychologists reacted in a similar way and were very impressed by significant results. It must be outlined that professional applied statisticians were not immune to misinterpretations, especially in the case of nonsignificance. However, the interpretations that accustomed users attach to the outcome of NHST can vary from one individual to another, and it is hard to conceive that there could be a consensus in the face of seemingly conflicting situations. In fact, beyond the superficial report of "erroneous" interpretations, it can be seen in the misuses of NHST intuitive judgmental "adjustments" that try to overcome its inherent shortcomings. These findings encourage the many recent attempts to improve the habitual ways of analyzing and reporting experimental data.

Nous avons étudié la manière dont des utilisateurs expérimentés interprètent les résultats des Tests de Signification de l'Hypothèse Nulle. Une étude empirique a été menée pour comparer les réactions de deux populations d'utilisateurs, des chercheurs en psychologie et des statisticiens professionnels, face à des situations conflictuelles. On présentait aux sujets les résultats d'une expérience planifiée pour tester l'efficacité d'un médicament en comparant deux groupes (traitement/placebo). Quatre situations étaient construites en combinant l'issue du test *t* (significatif vs. non-significatif) et la différence observée *D* entre les deux moyennes (grande vs. petite). Deux de ces situations apparaissaient conflictuelles (*t* significatif/*D* petite et *t* non-significatif/*D* grande). Trois aspects fondamentaux de l'inférence statistique étaient examinés au moyen de questions ouvertes: tirer une conclusion inductive sur la grandeur de la vraie différence, faire une prédiction relative à des données futures et prendre une décision sur l'arrêt de l'expérience. Les sujets étaient 25 statisticiens de l'industrie pharmaceutique en France, donc experts en statistique, et 20 chercheurs en psychologie de différents laboratoires français,

Requests for reprints should be addressed to Marie-Paule Lecoutre, ERIS, Laboratoire Psy.co, EA 1780, Université de Rouen, UFR Psychologie, Sociologie, Sciences de l'Éducation, 76821 Mont-Saint-Aignan Cedex, France **(E-mail:** marie-paule.lecoutre@univ-rouen.fr).

http://www.tandf.co.uk/journals/pp/00207594.html DOI: 10.1080/00207590244000250

ayant tous une expérience de l'analyse des données expérimentales. Dans l'ensemble, les statisticiens et les psychologues se sont comportés d'une manière similaire et ont été très influencés par les résultats significatifs. Un résultat important est que les statisticiens ne sont pas à l'abri des abus d'interprétation des tests, en particulier quand le résultat est non significatif. Cependant l'interprétation des tests peut varier considérablement d'un individu à l'autre et est loin de donner lieu à un consensus face à des situations en apparence conflictuelles. En fait, au delà du constat superficiel de l'existence d'interprétations "erronées", on peut voir dans les mésusages des tests des "ajustements" de jugement intuitifs, pour tenter de surmonter leurs insuffisances fondamentales. Ces résultats encouragent les nombreuses tentatives récentes d'améliorer les procédures habituelles pour analyser les données expérimentales et présenter les résultats.

Investigamos la manera en la que usuarios experimentados interpretan los resultados de las Pruebas de Significacia de la Hipótesis Nula (PSHN). Se diseñó un estudio empírico para comparar las reacciones de dos poblaciones de usuarios de las PSHN, psicólogos investigadores y profesionales de la estadística aplicada, enfrentados a situaciones contradictorias. Los participantes del estudio se enfrentaron a los resultados de un experimento diseñado para someter a prueba la eficacia de un fármaco en el que se comparaban dos grupos (tratamiento/placebo). Se construyeron cuatro situaciones en las que se combinaba el resultado de la aplicación de la prueba *t* (significativo vs no significativo) y las diferencias observadas entre las dos medias *d* (grandes o pequeñas). Estas dos situaciones eran conflictivas (*t* significativa/ *D* pequeña, y *t* no significativa / *D* grande). Se investigó tres aspectos fundamentales de la inferencia estadística por medio de preguntas abiertas: derivación de conclusiones inductivas sobre la magnitud de la diferencia verdadera de los datos disponibles, realización de predicciones para datos futuros, y toma de decisiones sobre si dar por terminado el experimento. Los participantes fueron 25 profesionales de la estadística de compañías farmacéuticas en Francia, versados en estadística, y 20 psicólogos investigadores con experiencia en el procesamiento y análisis de datos experimentales. En total, los estadísticos y los psicólogos respondieron de manera similar y se mostraron impresionados por el hecho de que los resultados fuesen sugnificativos. Debe subrayarse que los profesionales de la estadísticas no eran inmunes a las malas interpretaciones, especialmente en el caso de la no significancia. No obstante, las interpretaciones que los usuarios habituados adjudican al resultado de las PSHN pueden variar de un individuo a otro, y es difícil concebir que hubiera consenso frente a situaciones ostensiblemente conflictivas. De hecho, más allá del informe superficial de las interpretaciones "erróneas", puede apreciarse el mal uso de los "ajustes" en el juicio intuitivo de las PSHN que intenta corregir sus limitaciones inherentes. Estos hallazgos promueven los muchos intentos recientes por mejorar las formas habituales de analizar e informar sobre los datos experimentales.

How do experienced users such as professional applied statisticians or scientific researchers use Null Hypotheses Significance Testing (NHST) outcomes to interpret experimental data? In spite of all the rhetoric that denounced its widespread misinterpretations, NHST still remains the most ubiquitous statistical inference procedure, even when confidence intervals, likelihood, or Bayesian methods are clearly more appropriate (e.g., Goodman & Berlin, 1994; Nester, 1996; Rouanet, 1996). The main reason for the inadequacy of NHST is not that it is an incorrect normative model, but rather that it does not address the questions that scientific research requires. If the test is statistically *significant*, the null hypothesis is rejected in favour of the alternative hypothesis. This provides *no information* about the departure from the null hypothesis. When the sample is large a descriptively small departure may be significant. If the test is *nonsignificant*, the null hypothesis cannot be rejected. However, *this is not evidence favouring the null hypothesis*. In particular, a descriptively large departure from the null hypothesis may be nonsignificant if the experiment is sufficiently insensitive. Thus, users must resort to a more or less "naive" mixture of NHST results and other information, in other words they must make "judgmental adjustments" (Bakan, 1966; Phillips, 1973, p. 334) that try to overcome the inherent shortcomings of NHST.

Several empirical studies have investigated how well psychology students and/or researchers interpret NHST ((Falk & Greenbaum, 1995; Gordon, 2001; Mittag & Thompson, 2000; Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Most of these studies have emphasized the widespread existence of common misinterpretations of NHST. Recently, Haller and Krauss (2001) found out that most methodology instructors who teach statistics to psychology students, including professors who work in the area of statistics, share their students' misinterpretations. However, it is often taken for granted that statisticians are "armed with an understanding of the limitations of traditional methods" and "interpret quantitative results, especially *p*-values, very differently from how most nonstatisticians do" (Goodman, 1999, p. 1003). Schmidt (1995) describes the tasks of statisticians in pharmaceutical companies in the following terms: "Actually, what an experienced statistician does when looking at *p*-values is *to combine* them with information on sample size, null hypothesis, test statistic, and so forth to form in his mind something that is pretty much like a confidence

TABLE 1
The four situations and their corresponding normative answers

Situation and appearance	t test[a]		Observed difference	Standard Bayesian probabilities			Normative answer
				Pr(δ<–3)	Pr(–3<δ<+3)	Pr(δ>+3)	
1 *Nonconflicting*	t=3.674	p=.001	D=6.07 (*large*)	<.001	.037	.963	*Clinically interesting effect*
2 *Nonconflicting*	t=0.683	p=.50	D=1.52 (*small*)	.026	.719	.256	*No firm conclusion*
3 *Conflicting*	t=3.674	p=.001	D=1.52 (*small*)	<.001	.999	.001	*No clinically interesting effect*
4 *Conflicting*	t=0.683	p=.50	D=6.07 (*large*)	.158	.208	.634	*No firm conclusion*

[a]With constant sample size, different t- and p-values for identical observed difference (D) result from different within-group variances.

interval to be able to interpret the p-values in a reasonable way" (p. 490, emphasis added). More specifically, taking into account a measure of effect size can effectively prevent NHST users from some misinterpretations, especially in the situations where a nonsignificant result is combined with a large observed effect size.

Given these specificities we designed an empirical study to investigate to what extent accustomed NHST users appropriately combine the various "ingredients" usually available in statistical analyses, most frequently the NHST outcome and a descriptive measure of effect size. The present study came within the scope of a research project aimed at describing and analyzing the practices and attitudes of scientific researchers with regard to statistical inference (B. Lecoutre, 1983; M.-P. Lecoutre, 2000; M.-P. Lecoutre & Rouanet, 1993; Poitevineau, 1998; Poitevineau & Lecoutre, 2001). It was specifically designed to compare the reactions of two populations of NHST users, psychological researchers and professional applied statisticians, when faced with contradictory situations. These situations induce an apparent *conflict* between the outcome of a usual t test and the associated observed difference between the two means. Three fundamental aspects of statistical inference were investigated by means of open questions: (1) drawing inductive conclusions about the magnitude of the true difference from the data in hand, (2) making predictions about future data, and (3) making decisions on whether to stop or continue collecting more data.

METHOD

Subjects

The subjects were 25 professional statisticians from pharmaceutical companies in France, subjects well versed in statistics, and 20 psychological researchers from various laboratories in France, all with experience in processing and analyzing experimental data.

Material

The subjects were presented with the results of a study designed to test the efficacy of a drug by comparing two groups (treatment vs. placebo) of 15 patients each. The following evaluation criterion for the efficiency of the drug was given to the subjects: the drug was to be considered clinically interesting by experts in the field, if the unstandardized difference between the treatment mean and the placebo mean was more than +3. Four situations (see Table 1) were constructed by crossing the outcome of the t test (significant vs. nonsignificant) and the unstandardized observed mean difference D (large vs. small). Two of these situations appeared as conflicting: t significant/D small (situation 3) and t nonsignificant/D large (situation 4).

Questions

The situations were simultaneously presented. The subjects were asked the following three questions. (1) For each of the four situations, what conclusion would you draw for the efficacy of the drug? Justify your answer. (2) Initially, the experiment was planned with 30 subjects in each group and the results presented here are in fact intermediate results. What would be your prediction of the final results for D then t, then for the conclusion about the efficacy of the drug? (3) From an economical viewpoint, it would of course be interesting to stop the experiment with only the first 15 subjects in each group. For which of the four situations would you make the decision to stop the experiment, and conclude? Justify your answer.

The subjects were requested to respond in a spontaneous fashion, without making explicit calculations, and it was stressed that the task was an investigation of their statistical practices rather than a test of their theoretical knowledge. The responses were gathered individually and were completed by semidirective interviews aiming to compile further comments and justifications. The duration ranged in length from 15 to 25 minutes.

TABLE 2
Predictive probabilities for the final result of the experiment (60 subjects) given the intermediate results (30 subjects) for the four situations[a]

Situation and appearance	*t test*		*Observed difference*	*Standard predictive probabilities*		
				Pr(D>+3)	*Pr(t>+1.672)*	*Pr(l>+3)*
1 *Nonconflicting*	t=3.674	p=.001	D=6.07 (*large*)	.993	.998	.819
2 *Nonconflicting*	t=0.683	p=.50	D=1.52 (*small*)	.177	.244	.005
3 *Conflicting*	t=3.674	p=.001	D=1.52 (*small*)	<.001	.998	<.001
4 *Conflicting*	t=0.683	p=.50	D=6.07 (*large*)	.686	.244	.120

[a]D: observed difference; t: test statistic; $\underline{l}$: 95% lower confidence (or credible) limit

Normative answers

Concerning the efficacy of the drug

From a normative viewpoint, the task involves the following simple and general result: The $100(1-\alpha)\%$ interval estimate for the true difference δ can be computed from the t statistic and the observed difference D as $D \pm (D/t)t_{1-\alpha/2}$ (if $D \neq 0$), where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$ percentile of the Student's t distribution with the appropriate degrees of freedom (here 28). Similarly, the standard or fiducial Bayesian posterior distribution of δ (based on a noninformative prior distribution) is a generalized t distribution, centred on D and with scale factor D/t (B. Lecoutre, 1985). In other words, $(D/t)^2$ is an estimate of the sampling error variance; hence, for a given observed difference D, the higher the t statistic, i.e., the smaller the p-value, the closer δ and D must be. These straightforward and easily interpretable results should theoretically prevent the two main erroneous interpretations of NHST, which consist of (1) confusing statistical significance with substantive or scientific significance (see, e.g., Boring, 1919; Carver, 1978; Cox, 1977) and (2) interpreting a nonsignificant result as proof of the null hypothesis (e.g., Finch, Cumming, & Thomason, 2001; Harcum, 1990).

Thus, from a normative viewpoint, situations 1 and 3 are "favourable" in the sense they lead to the respective conclusions "clinically interesting effect" and "no clinically interesting effect." On the contrary, situations 2 and 4 cannot lead to firm conclusions, because of the great variability observed, even if the implications do still largely differ between these two situations. These normative answers can be legitimated by standard Bayesian statements (see Table 1), as well as by confidence intervals.

Concerning the prediction for the final result

In situations like the present one, where there is no available information other than the data, the standard Bayesian methods yield *predictive probabilities* that can be taken as reference probabilities for predicting the final results. For each of the four situations, Table 2 gives the predictive probabilities of obtaining, at the planned end of the experiment: an observed difference greater than +3 ($D > +3$); a significant t test at one-sided level $\alpha = .05$ ($t > +1.672$, 58 *df*); and a 95% lower confidence (or credible) limit greater than +3 ($\underline{l} > +3$). Theoretical results are given in Lecoutre (1999, 2001). A simple approximation method is given in the Appendix.

For situations 3 and 4, it is very unlikely that the conclusion of efficacy be asserted at the end of the experiment. For situation 3, this improbability might reinforce the decision to stop the experiment, as suggested by the conclusion that there is no clinically interesting effect obtained for the intermediate results. It is enlightening to contrast this result with the very high probability of the significant intermediate result being confirmed with additional data.

RESULTS

Results are summarized in Table 3, where the responses were coded into broad categories. It must be emphasized that all subjects perceived the task as routine for their professional activities. No subject stated that they would have liked to have additional information (such as standard deviations), except one statistician who stated that he would need confidence intervals to conclude. Furthermore, no subject suspected that the usual requirements underlying the t test (normality, equality of variances) could be violated (it was implicitly assumed that these requirements were fulfilled).

Question 1 (conclusion) and Question 3 (decision on stopping)

Situation 1 (significant test, large D, nonconflicting situation)

All subjects but one concluded that the drug was efficacious. Clearly, this was a consensual situation that was

TABLE 3
Responses (in percentiles) to Questions 1 to 3 for the four situations and the two subject groups[a]

		Situation 1 Sig *t*, large *D* *Nonconflicting* *Clinically interesting effect*		Situation 2 Nonsig *t*, small *D* *Nonconflicting* *No firm conclusion*		Situation 3 Sig *t*, small *D* *Conflicting* *No clinically interesting effect*		Situation 4 Nonsig *t*, large *D* *Conflicting* *No firm conclusion*	
	Response	*Stat*	*Psy*	*Stat*	*Psy*	*Stat*	*Psy*	*Stat*	*Psy*
Question 1:	Efficacy	96%	100%	0	0	*12%* [1]	*45%*	12%	0
Conclusion	Inefficacy	0	0	84%	85%	*80%* [2]	*40%*	36%	35%
	Do not know	4%	0	16%	15%	8%	15%	52%	65%
Question 2:	The same	68%	70%	52%	55%	64%	55%	52%	50%
Prediction	Increasing	0	10%	0	5%	0	5%	0	10%
about D	Decreasing	0	0	4%	0	4%	5%	0	0
	Do not know	32%	20%	44%	40%	32%	35%	48%	40%
Question 2:	The same	52%	60%	*16%* [3]	*40%*	48%	50%	16%	35%
Prediction	Increasing	8%	15%	16%	10%	8%	10%	28%	30%
about t	Decreasing	0	5%	0	5%	0	5%	0	5%
	Do not know	40%	20%	68%	45%	44%	35%	56%	30%
Question 2:	The same	84%	75%	52%	60%	76%	60%	28%	45%
Prediction	Efficacy	0	0	0	0	4%	0	20%	20%
about	Inefficacy	0	0	8%	0	4%	5%	0	0
efficacy	Do not know	16%	25%	40%	40%	16%	35%	52%	35%
Question 3:	Stopping	88%	75%	52%	55%	52%	60%	*4%* [4]	*30%*
Decision	Continuing	12%	25%	44%	40%	48%	40%	*96%* [5]	*60%*
	Do not know	0	0	4%	5%	0	0	0	10%

[a]*Stat* = statisticians, $n = 25$, and *Psy* = psychologists, $n = 20$. Italic characters indicate main differences between the two groups (all significant at level .05, Fisher's conditional test). The magnitude of the difference between the two parent proportions φ_p and φ_s can be assessed with a standard Bayesian procedure (Lecoutre, Derzko, & Grouin, 1995). Each of the following statements holds with probability .90: [1]$\varphi_p - \varphi_s > 0.16$; [2]$\varphi_s - \varphi_p > 0.22$; [3]$\varphi_p - \varphi_s > 0.07$; [4]$\varphi_p - \varphi_s > 0.12$; [5]$\varphi_s - \varphi_p > 0.21$.

considered to be particularly easy and favourable. For Question 3, responses were also clear-cut, with a high majority for stopping (88% among statisticians and 75% among psychologists). It should be noted that some psychologists answered this question in a "scientific" way rather than in the economical way specified in the instructions. They gave typical comments such as "in this situation I'll continue because we've got something interesting."

Situation 2 (nonsignificant test, small D, nonconflicting situation)

This situation also gave rise to a considerable consensus for Question 1: 84% of the statisticians and 85% of the psychologists concluded inefficacy. Here was a demonstration of a nonsignificant result abusively interpreted as proof of having no effect. However, the subjects who concluded inefficacy were divided for Question 3. A little more than half of them perceived the situation as very favourable and decided to stop (57% and 53% respectively in the two groups). On the contrary, other subjects expressed their uncertainty about this conclusion by commenting, "this can change" or "one must see if the tendency is confirmed or invalidated."

Situation 3 (significant test, small D, conflicting situation)

This situation revealed differences between the two groups. The statisticians were relatively homogeneous and 80% of them concluded inefficacy, correctly taking into account the smallness of *D*. On the other hand, the psychologists were divided. Almost half of them (45%) concluded the efficacy of the drug, relying exclusively on the significant test and confusing "statistical significance" with "substantive significance." This attitude could be extremely strong: One subject stated that "experts are wrong, they must revise their criterion," while the other subjects explicitly acknowledged that they discarded the criterion. Almost as many psychologists (40%) concluded the inefficacy of the drug because of the smallness of the observed difference. They admitted a non-null effect (as the test was significant) but with a size too small to be clinically relevant, which was in harmony with the normative response. For Question 3

the two groups were relatively close. The decision to continue was as follows: 50% of the subjects who concluded efficacy and 46% of the subjects who concluded inefficacy decided to continue, thereby expressing the uncertainty concerning their current conclusion.

Situation 4 (nonsignificant test, large D, conflicting situation)

This situation was considered as conflicting by a majority (65% of the psychologists and 52% of the statisticians), who did not give a conclusion. However, it must be stressed that the test had such an impact that one third of the subjects (35% and 36%, respectively) erroneously concluded inefficacy, in spite of the large observed difference. Nevertheless, in this case Question 3 distinguished these subjects: Only one statistician (11%), as opposed to a majority of psychologists (57%), decided to stop the experiment, showing great confidence in their erroneous conclusion.

Question 2 (prediction)

For each of the four situations, when predicting the final result, the subjects of both groups essentially answered either "about the same," which was the majority response, or "I have no idea"; "I can predict nothing." The subjects perceived this question as very difficult. Most psychologists stated that "they were not familiar with this type of question," while they also recognized that it was a "relevant and important problem." They were particularly hesitant about the *t* test statistic and were more concerned with the *p* value, arguing that the role of sample size was easier to evaluate. The statisticians, although more accustomed to interim analyses, were also very hesitant. If the subjects understood the question to be asking a point estimate, the response "about the same" looks normatively reasonable for *D* to the extent that no information other than the data was given. But this response is no longer compatible for *t* since it does not take into account the increase in sample size. It was only for the conflicting situation 4 that there was a non-negligible rate of responses (respectively 30% and 28% in the two groups), indicating that the *p* level should decrease and "perhaps reach significance."

DISCUSSION

At the final analysis, predictions about the test statistic and the conclusion were generally not available or inconsistent. These findings suggest two possible interpretations: either our subjects had no representation of the fact that for a given observed difference *D*, the higher the *t* statistic the closer δ and *D* must be, or they used inappropriate heuristics, such as the "representativeness heuristic," according to which the subjective probability of an event or a sample is determined by the degree to which it is similar in essential characteristics to its parent population. This heuristic leads to various predictable and systematic errors; in particular, since sample size does not represent any properties of the population, it is expected to have little or no effect on judgment of likelihood (Kahneman, Slovic, & Tversky, 1982; Kahneman & Tversky, 1972). The subjects should have a high degree of confidence that any two samples from the same population must resemble each other, so that they could dismiss the role of sample size. These findings support Freeman's (1993) conjecture that "even statisticians seem to have very little idea of how the interpretation of *p*-values should depend of sample size."

On the whole, statisticians and psychologists reacted in a similar way and were very impressed by significant results. It must be outlined that professional statisticians were not immune to misinterpretations, especially in the case of nonsignificance. Contrary to what Goodman's (1999) and Schmidt's (1995) assertions could lead one to think, it is not actually an easy task, even for professional statisticians, to interpret *p* values "in a reasonable way." Most notably, in the case of nonsignificance, the larger part of the subjects appeared unable to properly combine the observed difference with the traditional *t* test of the no difference null hypothesis.

The common practice in experimental publications is to dichotomize each result (significant/nonsignificant) according to the NHST outcome. It can be hypothesized that this practice reflects a circumstantial attitude ("it's the norm"), meaning a "mechanical behavior" (Gigerenzer, 1991), or a socially approved "automatic routine" (Falk & Greenbaum, 1995). This attitude is reinforced by a natural cognitive tendency to "take a position" when being published, and in some way to arrange every NHST outcome in a "cognitive filing cabinet" where a significant test is filed under "there is an effect" and a nonsignificant test is improperly filed under "there is no effect" (see the significance hypothesis of Oakes, 1986). On the contrary, in our experiment, only a minority of subjects had a systematically clearcut attitude. Most subjects tried to qualify the interpretation of the significance test in relation to the observed difference, or showed uncertainty in their conclusion when they were asked about stopping the experiment. Thus, even in the current context of the *dictatorship* of significant results in publications, the interpretations that accustomed users attach to the outcome of null hypothesis significance

tests can vary from one individual to another, and it is hard to conceive that there could be a consensus in the face of seemingly conflicting situations.

Some of our results could be interpreted as an individual's lack of mastery. However, this explanation is hardly applicable to professional statisticians. It is more likely that these results reveal the fundamental inadequacy of NHST to the true needs of the users. More than a third of the psychologists in our experiment explicitly stated that they were dissatisfied with NSHT and expressed their need for inferential methods that would better fit their spontaneous data interpretations. These findings encourage the many recent attempts to improve the habitual ways of analyzing and reporting experimental data. Concrete proposals can be found in recent papers dedicated to psychologists (see, e.g., Brandstätter, 1999; Cumming & Finch, 2001; Frick, 1995; Jones & Tukey, 2000; B. Lecoutre & Derzko, 2001; B. Lecoutre & Poitevineau, 2000; Loftus & Masson, 1994; Richardson, 1996; Rogers, Howard, & Vessey, 1993; Rouanet, 1996; Schmidt, 1996). They are supported more and more by editorial policies that require authors to report effect size indicators and their confidence intervals, in addition to or in place of NHST (see, e.g., American Psychological Association, 2001; Heldref Foundation, 1997; Loftus, 1993; Murphy, 1997; Snyder, 2000; Thompson, 1994, 1996; Wilkinson and Task Force on Statistical Inference, 1999). In any case, this is a crucial time because we are in the process of defining new publication norms, which should create a shift of emphasis in the presentation and interpretation of experimental results. We argue in other publications that the Bayesian methods are ideally suited for this purpose (see B. Lecoutre, Lecoutre & Grouin, 2001; B. Lecoutre, Lecoutre & Poitevineau, 2001; Rouanet, Bernard, Bert, Lecoutre, Lecoutre, & Le Roux, 2000). Furthermore, these methods provide insightful interpretations of many common procedures, including *p* values and confidence intervals, in intuitively appealing and readily interpretable forms.

Manuscript received December 2001
Manuscript accepted November 2002

REFERENCES

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

Bakan, D. (1966). The test of significance in psychological research. *On method* (pp. 1–29). San Francisco: Jossey-Bass. Reprinted in D. E. Morrison & R. E. Henkel (Eds.), (1970), *The Significance test controversy—A reader* (pp. 231–251). London: Butterwoths.

Boring, E. G. (1919). Mathematical versus scientific significance. *Psychological Bulletin*, *16*, 335–338.

Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods in Psychological Research*, *4*, 33–46. Retrieved October 2002 from http://www. mpr-online.de

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378–399.

Cox, D. R. (1977). The role of significance tests (With discussion). *Scandinavian Journal of Statistics, 4*, 49–70.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532–575.

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

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

Freeman, P. R. (1993). The role of *p*-values in analysing trial results. *Statistics in Medicine, 12*, 1443–1452.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory and Cognition, 23*, 132–138.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "Heuristics and Biases". *European Review of Social Psychology, 2*, 83–115.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine, 130*, 995–1004.

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121*, 200–206.

Gordon, H. R. D. (2001). American Vocational Education Research Association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Education Research, 26*(2). Retrieved October 2002 from http://scholar.lib.vt.edu/ejournals/JVER/v26n2/gordon.html

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*(1). Retrieved October 2002 from http://www.mpr-online.de

Harcum, E. R. (1990). Methodological versus empirical literature: Two views on casual acceptance of the null hypothesis. *American Psychologist, 45*, 404–405.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95–96.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*, 411–414.

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

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

Lecoutre, B. (1985). How to derive Bayes-fiducial conclusions from usual significance tests. *Cahiers de Psychologie Cognitive, 5*, 553–563.

Lecoutre, B. (1999). Two useful distributions for Bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference, 77*, 93–105.

Lecoutre, B. (2001). Bayesian predictive procedures for designing and monitoring experiments. In *Bayesian Methods with Applications to Science, Policy and Official Statistics* (pp. 301–310). Luxembourg: Office for Official Publications of the European Communities.

Lecoutre, B., & Derzko, G. (2001). Asserting the smallness of effects in ANOVA. *Methods of Psychological Research, 6*(1), 1–32. Retrieved October 2002 from http://www.mpr-online.de

Lecoutre, B., Derzko, G., & Grouin, J.-M. (1995). Bayesian predictive approach for inference about proportions. *Statistics in Medicine, 14*, 1057–1063.

Lecoutre, B., Lecoutre, M.-P., & Grouin, J.-M. (2001). A challenge for statistical instructors: Teaching Bayesian inference without discarding the "official" significance tests. In *Bayesian methods with applications to science, policy and official statistics* (pp. 311–320), Luxembourg: Office for Official Publications of the European Communities.

Lecoutre, B., Lecoutre, M.-P., & 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.

Lecoutre, B.,, & Poitevineau, J. (2000). Aller au delà des tests de signification traditionnels: Vers de nouvelles normes de publication. *L'Année Psychologique, 100*, 683–713.

Lecoutre, M.-P. (1983). La démarche du chercheur en psychologie dans des situations d'analyse statistique de données expérimentales. *Journal de Psychologie Normale et Pathologique, 3*, 275–295.

Lecoutre, M.-P. (2000). And... what about the researcher's point of view? In H. Rouanet, J.-M. Bernard, M.,C. Bert, B. Lecoutre, M.-P. Lecoutre, & B. Le Roux, *New ways in statistical methodology: From significance tests to Bayesian inference* (2nd ed., pp. 65–95). Bern, Switzerland: Peter Lang.

Lecoutre, M.-P., & Rouanet, H. (1993). Predictive judgments in situations of statistical analysis. *Organizational Behavior and Human Decision Processes, 54*, 45–56.

Loftus, G. R. (1993). Editorial comment. *Memory and Cognition*, *21*, 1–3.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review, 1*, 476–490.

Mittag, K. C., & Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*, 14–20.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82*, 3–5.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41*, 1299–1301.

Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics, 45*, 401–410.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. New York: Wiley.

Phillips, L. D. (1973). *Bayesian statistics for social scientists*. London: Nelson.

Poitevineau, J. (1998). *Méthodologie de l'analyse des données expérimentales: Étude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, prescriptive et descriptive*. Unpublished PhD, Université de Rouen, France.

Poitevineau, J., & Lecoutre, B. (2001). The interpretation of significance levels by psychological researchers: The .05-cliff effect may be overstated. *Psychonomic Bulletin and Review, 8*, 847–850.

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments and Computers, 28*, 12–22.

Rogers, J. L., Howard, K. I., & Vessey, J. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553–565.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55*, 33–38.

Rouanet, H. (1996). Bayesian procedures for assessing importance of effects. *Psychological Bulletin, 119*, 149–158.

Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., & Le Roux, B. (2000). *New ways in statistical methodology: From significance tests to Bayesian inference* (2nd edition). Bern, Switzerland: Peter Lang.

Schmidt, K. (1995). Statistical tests and estimations (Background paper). *Drug Information Journal, 29*, 483–491.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115–129.

Snyder, P. (2000). Guidelines for reporting results of group quantitative investigations. *Journal of Early Intervention*, *23*, 145–150.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837–847.

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

Wilkinson, L., and Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Zuckerman, M., Hodgins, H., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4*, 49–53.

APPENDIX

Predictive probabilities for the final results

The notations are summarized in the table. ε^2 denotes

TABLE FOR APPENDIX

	Observed difference	*Sampling error variance*	*t test statistic*	*95% lower confidence limit*
First part (2×15 subjects)	D_1	ε^2	t_1	l_1
Second part (2×15 subjects)	D_2	ε^2	t_2	l_2
Whole experiment (2×30 subjects)	$D=(D_1+D_2)/2$	$\varepsilon^2/2$	t	l

the sampling error variance of the observed difference within each part of the experiment: $\varepsilon^2 = (2/15)\sigma^2$, where σ^2 is the within-group variance. Simple approximations can be computed by assuming a known value for ε^2. Then the sampling distribution of D_2 is $N(\delta,\varepsilon^2)$ and, given the data of the first part, the standard Bayesian distribution of δ is $N(D_1,\varepsilon^2)$. Consequently the standard predictive distribution of D_2 is $N(D_1,2\varepsilon^2)$: Intuitively the sampling uncertainty about the difference D_2 within the second part is added to the uncertainty about δ given the data of the first part. The required predictive probabilities for the final results of the experiment can be approximated from the distribution of D_2, using these equivalences:

$D > 3 \Leftrightarrow D_2 > 6 - D_1$, deduced from $D = (D_1 + D_2)/2$

$t > 1.672 \Leftrightarrow D_2 > 1.672(\varepsilon\sqrt{2}) - D_1$, from $t = (D_1 + D_2)/(\varepsilon\sqrt{2})$

$l > 3 \Leftrightarrow D_2 > 6 - D_1 + 1.672\varepsilon\sqrt{2}$, from $l = (D_1 + D_2)/2 - 1.672(\varepsilon/\sqrt{2})$

and estimating ε by D_1/t_1. Better approximations are obtained by replacing the normal distribution with the generalized t distribution (with 28 degrees of freedom). This gives exact probabilities for $D > 3$ and very good approximations for $t > 1.672$ (respectively for each of the four situations: 0.999, 0.243, 0.999, and 0.243) and for $l > 3$ (respectively: 0.826, 0.007, <0.001, and 0.123).