Killeen's Probability of Replication and Predictive Probabilities: How to Compute, Use, and Interpret Them

Bruno Lecoutre, Marie-Paule Lecoutre, Jacques Poitevineau

To appear in *Psychological Methods*

Killeen's Probability of Replication and Predictive Probabilities: How to Compute, Use and Interpret Them

Bruno Lecoutre[1], Marie-Paule Lecoutre[2], Jacques Poitevineau[3]

[1]Centre National de la Recherche Scientifique and Université de Rouen
ERIS, Laboratoire de Mathématiques Raphaël Salem, UMR 6085
Avenue de l'Université, BP 12
F 76801 Saint-Etienne-du-Rouvray (France)
E-mail: bruno.lecoutre@univ-rouen.fr

[2]Université de Rouen
ERIS, PSY-NCA, EA 4306
UFR de Psychologie, Sociologie et Sciences de l'éducation
F 76821 Mont Saint Aignan Cedex (France)
E-mail: marie-paule.lecoutre@univ-rouen.fr

[3] Centre National de la Recherche Scientifique and Université Paris 6
ERIS, IJLRA/LAM-LCPE, UMR 7190
11 rue de Lourmel
F 75015 Paris (France)
E-mail: jacques.poitevineau@upmc.fr

April 8, 2009

Address correspondence to : Bruno Lecoutre, ERIS, Laboratoire de Mathématiques Raphaël Salem, UMR 6085 CNRS and Université de Rouen, Avenue de l'Université, BP 12, F 76801 Saint-Etienne-du-Rouvray (France); e-mail: bruno.lecoutre@univ-rouen.fr.

Running head: predictive probabilities

ABSTRACT

Killeen's (2005a) probability of replication (p_{rep}) of an experimental result is the fiducial Bayesian predictive probability of finding a same-sign effect in a replication of an experiment. p_{rep} is now routinely reported in Psychological Science and has also begun to appear in other journals. However, there is little concrete, practical guidance for use of p_{rep} and the procedure has not received the scrutiny that it deserves. Furthermore, only a solution that assumes a known variance has been implemented. A practical problem with p_{rep} is identified: in many articles p_{rep} appears to be incorrectly computed, due to the confusion between one-tailed and two-tailed *p*-values. Experimental findings reveal the risk of misinterpretations of p_{rep} as the predictive probability of finding a same-sign *and significant* effect in a replication (p_{srep}). Conceptual and practical guidelines are given to avoid these pitfalls. They include the extension to the case of unknown variance. Moreover, other uses of fiducial Bayesian predictive probabilities, for analyzing, designing ("how many subjects?") and monitoring ("when to stop?") experiments, are presented. Concluding remarks emphasize the role of predictive procedures in statistical methodology.

Keywords: Bayesian inference, fiducial inference, Killeen's $p_{rep,}$ *p*-values, predictive probabilities.

Acknowledgments

We thank Peter Killeen, Geoff Cumming and Eric-Jan Wagenmakers for useful feedback. We are especially grateful to Scott Maxwell who encouraged the extension of an earlier version of this article and provided helpful suggestions and comments that have improved the final version.

"*The essence of science is replication: a scientist should always be concerned about what would happen if he or another scientist were to repeat his experiment*" (Guttman, 1977).

Killeen (2005a) defined the probability of replication (p_{rep}) of an experimental result as the probability of finding a same-sign effect in a replication of an experiment (for an up-to-date discussion, see Killeen, 2008). p_{rep} now routinely appears in Psychological Science. Moreover, it begins to be occasionnaly reported beyond Psychological Science. A Web of Science review (April 24, 2008) for articles citing Killeen (2005a) revealed that it was occasionally reported in at least 15 other journals in various fields. p_{rep} is essentially used in the analysis of contrasts between means. It is associated either with a Student's *t* test (with a *z* test in some rare cases) or an ANOVA *F* test with one degree of freedom (*df*) in the numerator. So we will restrict our attention to this situation. Most of our points will be illustrated by means of a numerical example.

Frequentist and Bayesian Probabilities

Statistical inference is concerned with both known quantities – the observed data – and unknown quantities – the parameters and the data that have not been observed. In the frequentist conception traditionally perpetuated by the use of significance tests and confidence intervals, all the probabilities (in particular *p*-values and confidence levels) are *sampling* probabilities – that is frequencies – involving repetitions of the observations. They are always conditional on fixed parameter values. On the contrary, in the Bayesian conception, parameters can also be probabilized. This results in distributions of probabilities that express our uncertainty about parameters – before observations (*prior* probabilities) and after observations (*posterior* probabilities conditional on data) – and about future data (*predictive* probabilities: prior predictive probabilities before observations and posterior predictive probabilities conditional on data after observations).

So the *p*-value of the usual frequentist *t* test for comparing two means is the proportion of repeated samples for which the *t* statistic exceeds the (known) value observed in the data in hand, if the null hypothesis is true, that is if the true difference is assumed equal to zero. In the same way the frequentist confidence level, say 95%, for this true difference is based on the following universal statement: whatever value is assumed for the true difference, in 95% of the repeated samples the interval that should be computed includes this value.

On the contrary, p_{rep} is a probability conditional on the data in hand (and not on unknown quantities) and going to the unknown future observations (the replication). Consequently, p_{rep} is a Bayesian *predictive* expression of the statistical output of an experiment and breaks with the frequentist conception. It follows that for the first time a Bayesian probability is routinely reported in psychological journals.

Following Killeen (2005b), p_{rep} may be justified either from Fisher's fiducial argument or from a Bayesian argument assuming noninformative priors. Such priors are vague distributions that, *a priori*, do not favor any particular value. Consequently, they let the data ''speak for themselves'' (Box and Tiao, 1973, p. 2). We call this "noninformative" Bayesian approach *fiducial Bayesian* (Lecoutre B., 2000, 2008; Lecoutre, Lecoutre & Poitevineau, 2001), since it has the same incentive as Fisher's fiducial approach to only express evidence from data in terms of probability about parameters. Thus the fiducial Bayesian paradigm provides reference methods appropriate for situations involving scientific reporting. An alternative name could be

"objective Bayesian analysis", which was proposed by Berger (2004, p. 3) who clearly denounced "the common misconception that Bayesian analysis is a subjective theory."

Exact and broad replications

The usual frequentist definition of *p*-values and confidence levels involves imaginary 'exact' replications of the experiment with a fixed true difference. In this article, as usually done we will restrict our attention to these *conventional* replications. It must be acknowledged that practical replications are conducted "with incidental or deliberate changes to the experiment" (Cumming, 2008, p. 288). These *broad replications* involve different parameter values and require an average over the parameter space. Consequently, what is conceptually relevant in this case is a joint frequentist-Bayesian principle (Bayarri & Berger, 2004, p. 60). However, making predictions about broad replications would require an informative prior expressing some knowledge and assessing the changes made to the initial experiment, and presumably would not be easily accepted by frequentists.

Implementation and Practical Problems

For computing p_{rep} in practice, only a solution that assumes a known variance has been implemented. This implies that the *p*-value of the *t* test is considered as resulting from a *z* test. More than one hundred years after Student's famous article (Student, 1908), one can hardly be satisfied with this unnecessary (and generally unrealistic) assumption of known variance.

In any case, there is little concrete, practical guidance for computations, p_{rep} being not directly available from the standard statistical packages. Killeen (2005a, p. 353) suggested computing p_{rep} from the *p*-value. For the known variance case, he gave the exact following Excel formula:

$$p_{rep} = \text{NORMSDIST(NORMSINV(1-}p\text{)/SQRT(2))} \qquad (1)$$

and suggested the following approximation

$$p_{rep} \approx \left[1+\left(\frac{p}{1-p}\right)^{2/3}\right]^{-1}. \qquad (2)$$

Note that these formulae involve the *one-tailed p*-value.

The following example (called CC example) will serve us for illustration throughout the paper. Conway and Christiansen (2006) considered the mean of the n=10 paired differences between an experimental group and its baseline. For the auditory modality, they reported the results "$t(9) = 1.10$, $p = .30$, $p_{rep} = .76$, $d = .35$", where p is the two-tailed *p*-value and d is the standardized mean difference (Cohen's d). From their Table 1, it can be inferred that the unstandardized mean difference is 5.0. We will take 5.0 and $p = .30$ as the exact values for computations (hence $s = 14.3777\ldots$ and $t = 1.0997\ldots$). Note that we have deliberately chosen a nonsignificant example, because it is particularly important to alert readers to the fact that a nonsignificant result cannot be interpreted as "proof of no effect", or even as "proof of a small effect".

In this example, the function NORMSINV(1-p) in the Excel formula gives the *z* test statistic associated with the *p*-value (here $z = 1.036$), so that p_{rep} is the probability that a normal variable does not exceed $z/\sqrt{2} = .733$, hence a value slightly higher than the reported value: $p_{rep} = .768$ (the approximation gives the reported value .76). This assumes a known parent standard deviation $\sigma = 15.26$ which can be deduced from $s/\sigma = z/t = .942$.

The fact that Killeen's formulae (1) and (2) involve the *one-tailed p*-value is a possible source of confusion, although the author explicitly stated: "for two-tailed comparisons, halve *p*"

(Killeen, 2005a, p; 353). Indeed, a careful examination of the articles published in *Psychological Science* revealed that in many cases p_{rep} is incorrectly computed and that this seems due to the confusion between one-tailed and two-tailed *p*-values.

Our Work About Fiducial Bayesian Methods

Our previous work gives us tools that may contribute to analyze and improve statistical practice. For more than thirty years now, with other colleagues in France we have worked in order to develop routine Bayesian methods for the most familiar situations encountered in experimental data analysis. We have especially developed fiducial Bayesian methods based on noninformative priors (for an introduction, see Lecoutre, Lecoutre & Poitevineau, 2001; Lecoutre, 2006a, 2008). In particular, we studied the role of Bayesian predictive probabilities for designing ("how many subjects?") and monitoring ("when to stop?") experiments. For a long time in biostatistics, it has been recognized that "an essential aspect of the process of evaluating design strategies is the ability to calculate predictive probabilities of potential results" (Berry, 1991, p. 81). Bayesian predictive probabilities can answer essential questions such as: "how large a future experiment needs to be to have a reasonable chance to be in some sense conclusive?"; "given the data in hand at an interim stage of an experiment, what is the chance that the final result will be conclusive, or on the contrary inconclusive?"

Our Experimental Project About the Use of NHST

In close connection with this statistical work, an experimental project was conducted about the use of null hypothesis significance testing (NHST) by scientific researchers and applied statisticians (e.g., Lecoutre, Lecoutre & Poitevineau, 2001; Poitevineau & Lecoutre, 2001; Lecoutre, Poitevineau & Lecoutre, 2003). In particular, we investigated statistical prediction situations, which for instance consisted in asking subjects to estimate the probability, given a significant result, that this result would be significant once again in a replication of the experiment. A striking finding was that p_{rep}, the predictive probability of a same-sign result, could be confused with p_{srep}, the predictive probability of a same-sign *and significant* result, by at least one third of the subjects.

Organization of the Paper

The present article is divided into three parts. In the first descriptive part, we examine the current use of p_{rep} in Psychological Science and we report some experimental findings about statistical prediction situations that show the difficulties of interpretation of p_{rep}. In the second normative part, we develop general predictive procedures for contrasts between means. We first present Fisher's interpretation of the *p*-value as a predictive probability of replication under the null hypothesis. Then we consider the procedures for computing p_{rep} and p_{srep} in the usual case of unknown variance. Last, we develop other predictive procedures, about effect sizes and interval estimates. In the third prescriptive part, we examine the practical computation of predictive probabilities and their uses. Tables and computer programs are available for easy computations. Concluding remarks emphasize the role of predictive procedures in statistical methodology.

Part One: Descriptive Aspects

The Current Practice in Psychological Science

We analyzed the use of p_{rep} in all articles reporting statistical results in the October 2006 and October 2007 issues of Psychological Science. Results are summarized in Table 1. A large majority of articles (20 out of 27) reported p_{rep}. It replaced the *p*-value in 11 articles and was added to it in the 9 others. In 13 of these articles p_{rep} was reported only for significant results; in 8 of them nonsignificant tests were reported without p_{rep}, while in the 5 other articles no nonsignificant test was reported. This reflects an unfortunate difference of status between significant and nonsignificant results.

Table 1

*Use of p_{rep} in Psychological Science. Number of Articles Reporting it, in Addition (p_{rep}+*p*) to or in Place (Only p_{rep}) of the p-Value, Either for Significant* (S) *and Nonsignificant* (NS) *Tests, or Only for Significant Tests. The Right Column Is the Total Number of Articles Reporting p_{rep}.*

Articles reporting	October 2006 issue 15 articles		October 2007 issue 12 articles		27 articles
p_{rep} or *p*	p_{rep}+*p*	only p_{rep}	p_{rep}+*p*	only p_{rep}	p_{rep}
for S and NS tests*	3	2	1	1	7
only for S tests	1	3	2	2	8
for S tests**	1	2	1	1	5
	5	7	4	4	20

*except when $t<1$ or $F<1$ in two (2006) articles

**in these 5 articles no nonsignificant test was reported

In each article where p_{rep} was reported, there was at least one value associated with a *z* test, a *t* test or an *F* test for analyzing a contrast between means. For all the relevant cases[1], we computed p_{rep} from the test statistic and from the *p*-value (if given), by using both the exact (assuming a known variance) and approximate Killeen's formulae. The values reported in the articles were found to be in agreement with either the exact or the approximate formula in 10 articles[2]. In the other 10 articles, p_{rep} was systematically undervalued. In fact we found that for 8 of them, the values were in agreement with the formulae if we (erroneously) computed them with the two-tailed *p*-value. For the two remaining articles we were unable to get the reported values.

In summary, serious problems with the implementation of p_{rep} were revealed. In particular, in half of the articles the reported values appeared to be erroneous. In most cases, this can be explained by the confusion between one-tailed and two-tailed *p*-values. The authors who report p_{rep} merely add it to the test statistic and/or the *p*-value, in most cases without any reference and interpretation. Killeen's (2005a) article was mentioned in only 6 articles and only the following ambiguous definitions were found in 3 articles: "indicator of replicability" (Ackerman, Shapiro, Neuberg, Kenrick, Vaughn Becker, Griskevicius, Maner & Schaller, 2006, p. 838; Murphy, Steele

[1] In two articles a p_{rep} value was given for *F* ratios with several df in the numerator, leading to a puzzling interpretation. These cases were discarded. Note again that we found a few instances of p_{rep} values less than .5. They were associated to negative observed differences and the authors reported the probability of finding a positive difference in a replication, hence in fact 1-p_{rep}.

[2] In most articles reported values were not sufficiently accurate to distinguish between the two formulae.

& Gross, 2007, p. 882) and "the probability of replicating an effect of this size" (Dixon, Durrheim & Tredoux, 2007, p. 869). Consequently, we are not ensured that p_{rep} users, and their readers, correctly interpret it.

Some Experimental Findings About the Interpretation of p_{rep}

In an informal study, we recently asked psychological researchers to read one of the analyzed articles and to comment about p_{rep}. All of them stated that they were not previously informed of this practice, and a majority said that it was presumably the probability of finding again "the same" or "about the same result" in a replication, hence in the case of significance to find again a significant result. This fact that most people have a high degree of confidence that any two samples from the same population must resemble each other is well established (Kahneman, Slovic & Tversky, 1982). So, Cumming, Williams and Fidler (2004) investigated researchers' beliefs about the chance that a replication mean would fall within an original CI and concluded that "responses from 263 researchers suggest that many leading researchers in the 3 disciplines [psychology, behavioral neuroscience, and medicine] underestimate the extent to which future replications will vary" (p. 299). This fact is known as the "representativeness heuristic" (Kahneman & Tversky, 1972). The definition of p_{rep} "the probability of replicating an effect of this size" found in an article (Dixon, Durrheim & Tredoux, 2007, p. 869) was in accordance with this heuristic. In the same way an article published in another journal defined p_{rep} as "the probability that the experiment can be repeated and will give the same value" (Gerits, Van Belle & Moens, 2007, p. 4).

In an early study about the representativeness heuristic, Tversky and Kahneman (1971) asked researchers to estimate the probability, given a significant difference in a first experiment, of finding a significant result once again in a replication of the experiment. The probability values given as responses were found to be markedly higher than the reference Bayesian probability value (about .50). Oakes (1986) reported similar findings. This fact that most users of statistics overestimate the probability of replicating a significant result (p_{srep}) was outlined by Goodman (2002, p. 2446): "I have found experienced researchers, and even some statisticians, to be extremely surprised how low this replication probability actually is."

This first study was refined by Lecoutre and Rouanet (1993) who asked two different questions, one about the sign of the difference in the replication and the other about the result of the significance test (see also Lecoutre M.-P., 2000). Furthermore they considered various situations that differed according to the results obtained in the first experiment. In particular, in a situation similar to that reported by Tversky and Kahneman (1971), the researchers were given both descriptive (raw difference +1.82) and inferential ($t(19) = 2.09$, $p = .05$) results for the first experiment. The two questions were respectively: "what, for you, is the probability that in the replication the observed difference will be positive?" and "what, for you, is the probability that the observed difference will be positive, and the result of Student's t test will be at least significant?" They were asked to 50 psychological researchers, all with experience in processing and analyzing experimental data.

The answers were compared with the probabilities of replication[3]. For the first question this was precisely Killeen's p_{rep}, hence $p_{rep} = .92$. For the second question the probability of a same-

[3] It must be noted that the term "probability of replication" may be misleading in a sampling framework (Cumming, 2005), since it could be confused with the sampling probability of replication, conditional on the parameters. It may also be indefinite in a Bayesian framework, since it depends on a prior probability distribution (Macdonald, 2005).

sign significant replication (here at two-tailed level .05) p_{srep}, was about .50. These two values were computed without assuming a known variance, the way to do this being explained later in this paper. The striking finding was that half of the participants gave numerically close values for p_{rep} and p_{srep}. About one third of the subjects even gave exactly the same answer, and thus completely failed to distinguish p_{srep} from p_{rep}. Similar findings were observed in other situations involving different statistical results.

Lecoutre, Poitevineau and Lecoutre (2003) studied predictions about the final outcome of an experiment, given various intermediate statistical results obtained after half of the participants have been included. One of the questions asked to 25 professional statisticians from pharmaceutical companies in France and 20 psychological researchers was "what would be your prediction of the final results, firstly for the observed difference, and secondly then for the *t* test statistic?" The conclusion of the authors was again that most participants did not differentiate their predictions about the significance test from those about the observed difference.

Part Two: Normative Aspects

Fisher's Conception of the p-Value as a Predictive Probability

Killeen (2005a), as many others, referred to Fisher for the definition of the *p*-value. So it is enlightening to read the following definition given by Fisher (1990/1970).

> If *x* (for example the mean of a sample) is a value normally distributed about zero, and σ is its true standard error, then the probability that *x*/σ exceeds any specified value may be obtained from the appropriate table of the normal distribution; but if we do not know σ, but in its place have *s*, an estimate of the value of σ, the distribution required will be that of *s*/σ and this is not normal. The true value has been divided by a factor, *s*/σ, which introduces an error. […] the distribution of *s*/σ is calculable, and although σ is unknown, *we can use in its place the fiducial distribution of σ given s to find the probability of x exceeding a given multiple of s*. (p. 118) [italics added].

It is a definition of the probability that the standardized mean *x*/*s* (or equivalently the *t* statistic) exceeds any specified value, hence in particular the value observed in the data in hand: the so-called one-tailed *p*-value of the *t* test. In this definition, the observed standard deviation *s* is clearly treated as a fixed quantity and σ as a random variable. It is at odds with the commonly used frequentist conception, in which *s* is a random variable and σ a (unique) fixed quantity. Fisher used a fiducial argument to derive the posterior distribution of σ given *s*. So, for the CC example, we first consider the sampling probability (under H_0: μ = 0) that the *t* statistic exceeds 1.10 *for each possible value* of σ, *s being fixed* and equal to its observed value. This probability is given by the normal distribution centered on zero, with standard deviation σ/*s*. Some examples of values are given in Table 2. Then, to obtain the one-tailed *p*-value of the *t* test, these sampling probabilities are averaged over σ/*s* values. The weights are given by the fiducial Bayesian distribution of *s*/σ (where σ is the random variable and *s* is fixed), which is the same distribution as the sampling distribution of *s*/σ (where *s* is the random variable and σ is fixed),

---

Here a more precise term would be "fiducial Bayesian predictive probability of replication". "Probability of replication" is used as a shortcut for simplicity.

hence the square root of a chi-square distribution divided by its *df* (here 9)[4]. The resulting average value is .150, hence the two-tailed *p*-value .30.

Table 2
Fisher's Derivation of the p-Value as a Predictive Probability. For the CC Example, Probability that t Exceeds the Observed Value 1.10 as a Function of Given s/σ.

s/σ	Pr(t>1.10)
.482	.298
.773	.198
.904	.160
.942	**.150**
.963	.145
1.023	.130
1.166	.100
1.552	.044

Note. .942 is the known value assumed in the Killeen procedure; the other s/σ values are respectively the 1, 20, 40, 50, 60, 80 and 99 percent points of its fiducial Bayesian distribution. Pr(t>1.10) is the probability that the normal distribution centered on zero, with standard deviation σ/s, exceeds 1.10.

The result is the same as in the frequentist conception but the justification is quite different. Lecoutre (1985) called Fisher's conception a "semi-Bayesian significance test", since the Bayesian method is only applied to the *nuisance* parameter σ, the probability *p* being *conditional on the value of the true mean* μ *specified by the null hypothesis*. Bayesians have introduced the notion of *posterior predictive p*-value viewed "as the posterior mean of a classical [frequentist] *p* value, averaging over the posterior distribution of (nuisance) parameters under the null hypothesis." (Meng, 1994, p. 1142). This is precisely a Bayesian extension of Fisher's conception, even if this was not recognized by the Bayesian proponents. Without entering here into the frequentist/Bayesian debates, it must be stressed that Fisher "was in fact much closer to the 'objective Bayesian' position than that of the frequentist Neyman" (Zabell, 1992, p. 381).

Note again that posterior predictive *p*-values have been generalized beyond the case of a simple null hypothesis and are becoming a Bayesian standard for model checking (Gelman, Carlin, Stern & Rubin, 2004). They can be considered in the general framework of "measures of surprise" used to quantify "the degree of incompatibility of data with some hypothetized model H_0 without reference to any alternative model" (Bayarri & Berger, 1997, p. 1). Of course, since they are both conditional on H_0, from a methodological, not conceptual, viewpoint, the same pro and con arguments can be addressed to frequentist and to Bayesian posterior predictive *p*-values.

Derivation of p_{rep} and p_{srep} for Known σ

As is true of Fisher's *p*-value, Killeen's p_{rep} and p_{srep} are also defined as posterior predictive probabilities. Assuming σ is known, let z_{rep} denote the z test statistic in the replication and let z_{obs}

[4] This distribution can be justified in both the fiducial and Bayesian frameworks. In the latter case, the usual noninformative prior, uniform for log σ, is assumed.

be its observed value in the first experiment. The probability that z_{rep} exceeds any specified value is also an averaged sampling probability, but the roles of μ and σ are reversed, μ being the random variable and σ being assumed fixed (known).

p_{rep}, the probability of a same-sign effect in a replication, is the probability that z_{rep} exceeds 0 (assuming $z_{obs} > 0$). $p_{srep}(\alpha)$, the probability of a same-sign effect significant at two-tailed level α, is the probability that z_{rep} exceeds the critical value $z_{c(\alpha)}$, for instance 1.960 for $\alpha = .05$. We now consider the sampling probability that z_{rep} exceeds the specified value *for each possible value* of μ. This probability is given by the normal distribution centered on $\mu/(\sigma/\sqrt{n})$, with unit standard deviation. Some examples of values for the CC example are given in Table 3.

Table 3
Derivation of p_{rep} *and* $p_{srep}(.05)$ *for Known* σ*. For the CC Example,* $\Pr(z_{rep}>0)$ *Is the Sampling Probability of Finding a Same Sign Effect in a Replication and* $\Pr(z_{rep}>1.960)$ *Is the Sampling Probability of Finding a Same Sign and Significant at Two-Tailed Level .05 Effect.*

$\mu/(\sigma/\sqrt{n})$	$\Pr(z_{rep}>0)$	$\Pr(z_{rep}>1.960)$
-1.290	.099	.0006
.195	.577	.039
.783	.783	.120
1.036	**.850**	**.178**
1.290	.901	.251
1.878	.970	.467
3.363	.9996	.920

Note. $\Pr(z_{rep}>0)$ and $\Pr(z_{rep}>1.960)$ are given by the sampling distribution $N(\mu/(\sigma/\sqrt{n}),1)$, as a function of fixed $\mu/(\sigma/\sqrt{n})$. The $\mu/(\sigma/\sqrt{n})$ values are respectively the 1, 20, 40, 50, 60, 80 and 99 percent points of its fiducial-Bayesian distribution N(1.036,1) (1.036 being the z_{obs} value assumed in the Killeen procedure).

Then, to obtain p_{rep} and p_{srep}, these sampling probabilities are averaged over $\mu/(\sigma/\sqrt{n})$ values. The weights are given by the fiducial Bayesian distribution of $\mu/(\sigma/\sqrt{n})$, conditional on the observed value x_{obs}=5 and the known value σ assumed to be 15.26. It is a normal distribution centered on $z_{obs} = x_{obs}/(\sigma/\sqrt{n}) = 1.036$, with unit standard deviation[5]. The resulting average values are $p_{rep} = .768$ and $p_{srep}(.05) = .257$, given by the predictive distribution of z_{rep}, a normal distribution also centered on 1.036 and with standard deviation $\sqrt{2}$.

The fact that $p_{rep} = .768$ and $p_{srep}(.05) = .257$ are respectively markedly smaller and larger than the probabilities .850 and .178 associated with $\mu/(\sigma/\sqrt{n}) = 1.036$, the most likely value given the data in hand, can be intuitively understood. For p_{rep} for instance, Table 3 shows that two symmetrical values around 1.036, for instance .195 and 1.878, correspond to markedly asymmetrical probabilities around .850, respectively .577 and .970. Since .195 and 1.878 are equally likely, it results that the smaller probability .577 has the same weight as .970.

[5] This distribution can be justified in both the fiducial and Bayesian frameworks. In the latter case, the usual noninformative prior, uniform for μ, is assumed.

Derivation of p_{rep} and p_{srep} for Unknown σ: The K-Prime Distribution

The derivation of p_{rep} and p_{srep} when σ is unknown is a straightforward extension of the known σ case, the normal distribution being replaced with new distributions. Let t_{rep} denote the *t* test statistic in the replication and let t_{obs} be its observed value in the first experiment. In this case, we have to consider the probability that t_{rep} exceeds a specified value *for each possible value* of (μ,σ).

On the one hand, the sampling distribution of t_{rep} is a *noncentral t* distribution with ν *df* (as the *t* test) and noncentrality parameter $\mu/(\sigma/\sqrt{n})$, noted $t'_{(\nu)}(\mu/(\sigma/\sqrt{n}))$. This distribution is familiar to power analysts and confidence intervals users (see, e.g., Cumming & Finch, 2001; Lecoutre, 2007). On the other hand, the fiducial Bayesian distribution of $\mu/(\sigma/\sqrt{n})$ is a *lambda-prime* distribution with ν *df* and noncentrality parameter equal to the observed value $t_{obs} = 1.10$, noted $\Lambda'_{(\nu)}(t_{obs})$. This distribution, which was considered (with no name) by Fisher (1990/1973, pp. 126-127) in the fiducial framework, was called lambda-prime in Lecoutre (1999) (see also Rouanet & Lecoutre, 1983 and Lecoutre, 2007)[6]. Some examples of values for the CC example are given in Table 4. In this case we consider the critical value $t_{c(\nu,\alpha)}$, here 2.262 for $\nu = 9$ and $\alpha = .05$.

Table 4

Derivation of p_{rep} and $p_{srep}(.05)$ for Unknown σ. For the CC Example, $Pr(t_{rep}>0)$ *Is the Sampling Probability of Finding a Same Sign Effect in a Replication and* $Pr(t_{rep}>2.262)$ *Is the Sampling Probability of Finding a Same Sign and Significant at Two-Tailed Level .05 Effect.*

$\mu/(\sigma/\sqrt{n})$	$Pr(t_{rep}>0)$	$Pr(t_{rep}>2.262)$
-1.328	.092	.0008
.201	.580	.038
.808	.790	.108
1.069	**.957**	**.158**
1.331	.908	.221
1.938	.974	.410
3.474	.9997	.870

Note. $Pr(t_{rep}>0)$ and $Pr(t_{rep}>2.262)$ are given by the sampling distribution $t'_{(9)}(\mu/(\sigma/\sqrt{n}))$, as a function of fixed $\mu/(\sigma/\sqrt{n})$. The $\mu/(\sigma/\sqrt{n})$ values are respectively the 1, 20, 40, 50, 60, 80 and 99 percent points of its fiducial-Bayesian distribution $\Lambda'_{(9)}(1.10)$.

The resulting average values are $p_{rep} = .772$ and $p_{srep}(.05) = .230$. They are higher than the probabilities .768 and .257 obtained when σ is known. This is a consequence of the fact that for a given *p*-value t_{obs} is larger than z_{obs}, and thus is not really surprising. Formally, p_{rep} and p_{srep} are given by the predictive distribution of t_{rep}, called a *K-prime* distribution in Lecoutre (1984). By analogy with the normal distribution, we note

$$t_{rep} \mid t_{obs} \sim K'_{(\nu,\nu)}(t_{obs}, 2). \qquad (3)$$

[6] It can be justified in the Bayesian framework, assuming the usual noninformative prior, uniform for (μ,log σ).

The number of *df* ν is involved twice in the predictive distribution, since it is associated both with the sampling distribution and with the fiducial Bayesian distribution. Note that the *K-prime* distribution includes all the other involved distributions as particular cases (see Appendix A). In particular, for large values of ν the predictive distribution of t_{rep} is approximately normal, so that, as it can be expected, the predictive probabilities are close to the values computed assuming a known variance.

A simple general formula for p_{rep}. A remarkable property of the *K*-prime distribution is that the probability that it exceeds zero is given by the usual (central) Student *t* distribution. If we assume $t_{obs} > 0$, we get the general formula for a contrast between means:

$$p_{rep} = \Pr(t_{rep} > 0 \mid t_{obs}) = \Pr(t_{(\nu)} < t_{obs}/\sqrt{2}), \qquad (4)$$

A general formula for p_{srep}. Conceptually, it is no more difficult to get p_{srep}, the *t* distribution being replaced with the *K*-prime distribution:

$$p_{srep}(\alpha) = \Pr(t_{rep} > t_{c(\nu,\alpha)} \mid t_{obs}) = \Pr[K'_{(\nu,\nu)}(t_{obs}/\sqrt{2}) > t_{c(\nu,\alpha)}/\sqrt{2}], \qquad (5)$$

and more generally

$$\Pr(t_{rep} > T \mid t_{obs}) = \Pr[K'_{(\nu,\nu)}(t_{obs}/\sqrt{2}) > T/\sqrt{2}] \qquad (6)$$

where $K'_{(\nu,\nu)}(t_{obs}/\sqrt{2})$ is the standard (unscaled) *K*-prime distribution, with noncentrality parameter $t_{obs}/\sqrt{2}$.

Alternatively to p_{srep}, we can compute a $100(1-\alpha)\%$ credible[7] prediction interval for t_{rep}. The two limits of this interval are respectively $\sqrt{2}$ times the $100(\alpha/2)$ lower and upper percentiles of the $K'_{(\nu,\nu)}(t_{obs}/\sqrt{2})$ distribution.

Predictions About the Magnitude of a Contrast

In this section, we only consider predictions about Cohen's *d*, for reasons of simplicity. Indeed, since Cohen's *d* is proportional to the *t* test statistic, predictions about its value d_{rep} in a replication can be easily derived from predictions about t_{rep}. Procedures for unstandardized contrasts are given in Appendix B.

Predictions about Cohen's d. Using the fact that $d_{rep}/t_{rep} = d_{obs}/t_{obs}$ (assuming $d_{obs} \neq 0$), hence $d_{rep} = t_{rep}(d_{obs}/t_{obs})$, we get from (6) the predictive probability that d_{rep} exceeds *D*:

$$\Pr(d_{rep} > D) = \Pr[t_{rep} > (t_{obs}/d_{obs})D] = \Pr[K'_{(\nu,\nu)}(t_{obs}/\sqrt{2}) > (D/d_{obs})t_{obs}/\sqrt{2}], \qquad (7)$$

which is a general formula for a contrast between means. A prediction interval for d_{rep} can be deduced from the prediction interval for t_{rep}, multiplying the limits by d_{obs}/t_{obs}.

Predictions about the confidence limits of Cohen's d. Confidence intervals can be viewed as the set of hypothesized values that are nonsignificant. It follows that predictive probabilities about confidence limits for contrasts between means are also given by the *K*-prime distribution. The probability that the lower confidence limit for μ/σ in a replication exceeds a given value *L* is the predictive probability of a standardized effect larger than *L* and such that the null hypothesis $H_0 : \mu/\sigma = L$ is rejected at one-tailed level $\alpha/2$. This test is significant if t_{rep} exceeds the $100(\alpha/2)\%$ upper point of the noncentral *t* distribution with noncentrality parameter $L\sqrt{n}$, i.e. more generally for a standardized contrast $(t_{obs}/d_{obs})L$, denoted by $t'_{upp(\nu,\alpha/2)}[(t_{obs}/d_{obs})L]$. Note that

[7] Bayesians use "credible" instead of confidence to underline the difference in interpretation.

in this case we consider one-tailed level since the test involves a non-symmetrical distribution[8]. The predictive probability, given by (6), is:

$$\Pr[K'_{(\nu,\nu)}(t_{obs}/\sqrt{2}) > t'_{upp(\nu,\alpha/2)}((t_{obs}/d_{obs})L)/\sqrt{2}], \qquad (8)$$

Of course, for $L = 0$ it is $p_{srep}(\alpha)$ if $d_{obs} > 0$, and the probability of an opposite-sign significant replication if $d_{obs} < 0$.

We get in the same way the predictive probability that the upper confidence limit does not exceed U, the upper point being replaced with the $100(\alpha/2)\%$ lower point:

$$\Pr[K'_{(\nu,\nu)}(t_{obs}/\sqrt{2}) < t'_{low(\nu,\alpha/2)}((t_{obs}/d_{obs})U)/\sqrt{2}] \qquad (9)$$

Other Predictions

Predictions for a future experiment with different sample sizes. We can also derive the predictive probabilities for a future experiment with different sample sizes. If all the cell counts in the first experiment are multiplied by the same constant c (resulting in ν' *df* in the future experiment), all the above formulae can be easily generalized. For instance, the predictive distribution of t is

$$t_{future} \mid t_{obs} \sim K'_{(\nu,\nu')}(\sqrt{c}\, t_{obs}, 1+c). \qquad (10)$$

Note that c can be a fractional number and can be smaller than one, if this results in integer counts (which is not always the case, especially if cell counts are unequal).

In particular the probability of finding again a positive t value with counts multiplied by c is

$$\Pr(t_{(\nu)}) < t_{obs}\sqrt{c/(1+c)}, \qquad (11)$$

so that, when c tends to infinity, it tends to $\Pr(t_\nu < t_{obs}) = 1\text{-}p/2$. This result gives us another valuable interpretation of the one-tailed p-value as a predictive probability: it is the predictive probability of a different-sign result in a future extremely large data set. This fiducial Bayesian interpretation of the one-tailed p-value can be viewed as the counterpart of the Jones and Tukey (2000) frequentist view of NHST as a three alternative decision: the sign is positive, is negative, is not yet determined. Note again that when the cell counts tend to infinity the prediction intervals for standardized and unstandardized contrasts coincide with the confidence intervals for the corresponding parameters.

Predictions about future experimental units. For instance, when comparing two independent groups, it can be of interest to consider one future observation for each group and compute the predictive probability that their difference will be positive. This aspect would deserve more considerations, but is beyond the scope of this paper.

Predictions including the available data. An important question is to predict the final result of an experiment, given the available data at an interim stage. Let us suppose for the CC example that the experiment was planned with 20 subjects and that the data were obtained from an interim analysis after the inclusion of 10 subjects. The predictive distribution for the final unstandardized difference (given the interim data) is deduced from the predictive distribution of x_{rep} (since the final difference is $(x_{obs}+x_{rep})/2$). Other predictive distributions are generally not available in closed form but they can be easily obtained by simulation techniques. The principle is to simulate the fiducial Bayesian distribution associated with the available data and, for each generated value to simulate the sampling distribution of the future data.

---

[8] The notation $t'_{upp(\nu,\alpha/2)}$ makes explicit the fact that it is a one-tailed value, by contrast with the notation $t_{c(\nu,\alpha)}$ for the two-tailed critical value involved for the (symmetrical) usual t distribution.

Predictions about any event. From simulation techniques, we can get the predictive probability of any event of interest about the future data or the whole data.

Part Three: Prescriptive Aspects

How to Compute Predictive Probabilities?

User-friendly tools for easily computing p_{rep}, p_{srep}, and more generally the probabilities of replication involving the *K*-prime distribution, are available on our website http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris-Prep.html or upon request to the first author. They include tables, an Excel worksheet, a Word macro and computer programs. We give hereafter some numerical illustrations for the CC example.

Using Predictive Probabilities About Replications

What is the probability of a same-sign replication (p_{rep})? Killeen's Excel formula (1) for computing p_{rep} from the *one-tailed p*-value in the known variance case can be generalized to[9]

$$p_{rep} = \text{1-TDIST(TINV(2*p,df)/SQRT(2),df,1)} \qquad (12)$$

For more accuracy, and to avoid any confusion between two-tailed and one-tailed *p*-values, it is generally preferable to compute p_{rep} from the test statistic (Student's *t* or ANOVA *F* with one *df* in the numerator)

$$p_{rep} = \text{1-TDIST(ABS(t)/SQRT(2),df,1)} \quad \text{or} \quad p_{rep} = \text{1-TDIST(SQRT(F)/SQRT(2),df,1)} \qquad (13)$$

For the CC example, for $t = 1.10$ and $df = 9$, TDIST(ABS(t)/SQRT(2),df,1) gives $\Pr(t_{(9)} > 1.10/\sqrt{2}) = .228$, hence the p_{rep} value .772 previously obtained. Consequently, p_{rep} can be very easily computed in the unknown variance case, the normal distribution being simply replaced with the *t* distribution. This is coherent with the fact that the predictive distribution of the mean difference x_{rep} is a *t* distribution.

Table 5 extends Cumming's Table 1 for p_{rep} as a function of two-tailed *p*-value (Cumming, 2005, p. 1004) to some *df* values. The good news is that for low *df* the probabilities exceed Cumming's values, which are conservative: they underestimate the probability of replication. This is a consequence of the fact that for a given *p*-value the smaller *df*, the larger t_{obs} is.

Table 5

Predictive Probabilities of Replication p_{rep} and of Significant Replication at Two-Tailed Level .05 p_{srep}(.05) as a Function of Two-Tailed p-Value and Degrees of Freedom.

	p_{rep}					p_{srep}(.05)				
p/*df*	10	25	50	100	∞*	10	25	50	100	∞
.1.0	.500	.500	.500	.500	.500	.073	.079	.081	.082	.083
.50	.684	.684	.683	.683	.683	.161	.173	.178	.180	.182
.20	.823	.820	.819	.818	.818	.293	.307	.311	.313	.316
.10	.886	.881	.879	.878	.878	.397	.406	.409	.410	.412

[9] Note the difference in form between (1) and (12). While NORMSINV returns the cumulative distribution function, the TINV function returns the two-tailed probability. Consequently its argument is "2*p" (in fact the *two-tailed p*-value) instead of "1-p". Furthermore TDIST takes the final argument "1" to return the upper one-tailed probability.

.05	.927	.921	.919	.918	.917	.500	.500	.500	.500	.500
.02	.960	.954	.952	.951	.950	.626	.612	.607	.605	.602
.01	.976	.970	.968	.967	.966	.710	.685	.677	.672	.668
.005	.985	.980	.979	.977	.976	.782	.748	.737	.731	.725
.002	.992	.989	.987	.986	.986	.857	.817	.803	.795	.788
.001	.996	.993	.992	.991	.990	.899	.859	.843	.835	.827
.0001	.999	.998	.998	.997	.997	.974	.946	.931	.923	.914

*This column is the same as Killeen's p_{rep} assuming σ known.

What is the probability of a same-sign and significant replication (p_{srep})? The probability of a same-sign and significant replication $p_{srep}(.05)$ is reported in Table 5 for $\alpha = .05$. It is always substantially smaller than p_{rep}, so that the two values cannot, even approximately, be confused. For the CC example, we get from (5) $p_{srep}(.05) = \Pr(K'_{(9,9)}(1.10/\sqrt{2}) > 2.262/\sqrt{2}) = .230$. In spite of the nonsignificant result, on the basis of the data in hand there is a 23 per cent chance of finding a positive difference significant at two-tailed level .05 in a replication ($t_{rep} > 2.262$). There is also a 1.6 per cent chance of finding a negative and significant difference ($t_{rep} < -2.262$), hence a 24.6 per cent chance of finding a significant result ($|t_{rep}| > 2.262$).

What are prediction limits for t_{rep}? The 2.5 lower and upper percent points of the $K'_{(9,9)}(1.10/\sqrt{2})$ are -1.363 and 3.311. Multiplying these values by $\sqrt{2}$ we get the 95% prediction interval for t_{rep} [-1.93 , 4.68]. In spite of the nonsignificant result, this interval includes highly significant values (4.68 corresponds to a p-value of .001).

What is the probability of a small, medium, large Cohen's d_{rep}? Given a nonsignificant result, statistical users generally expect that in a replication Cohens' d will have a small, or at least limited value. Multiplying the above limits for t_{rep} by $d_{obs}/t_{obs} = .316$ $(1/\sqrt{10})$, we get the 95% prediction interval for d_{rep}. [-.61 , 1.48]. Alternatively, we can compute the predictive probability that $|d_{rep}|$ does not exceed a given value, for instance $\Pr(|d_{rep}|<.20) = .255$. Clearly, this contradicts the expectation of limited value.

Such predictive inference can help understanding the respective roles of sample sizes and p-values in statistical prediction. For illustration Figure 1 gives the predictive probabilities of finding $|d_{rep}|<.20$ as a function of two-tailed p-value and number of paired differences (n = 10, 50, 100, 1000). It is clear that a nonsignificant result cannot be interpreted as "a proof of a small difference", unless a very large sample size is used.

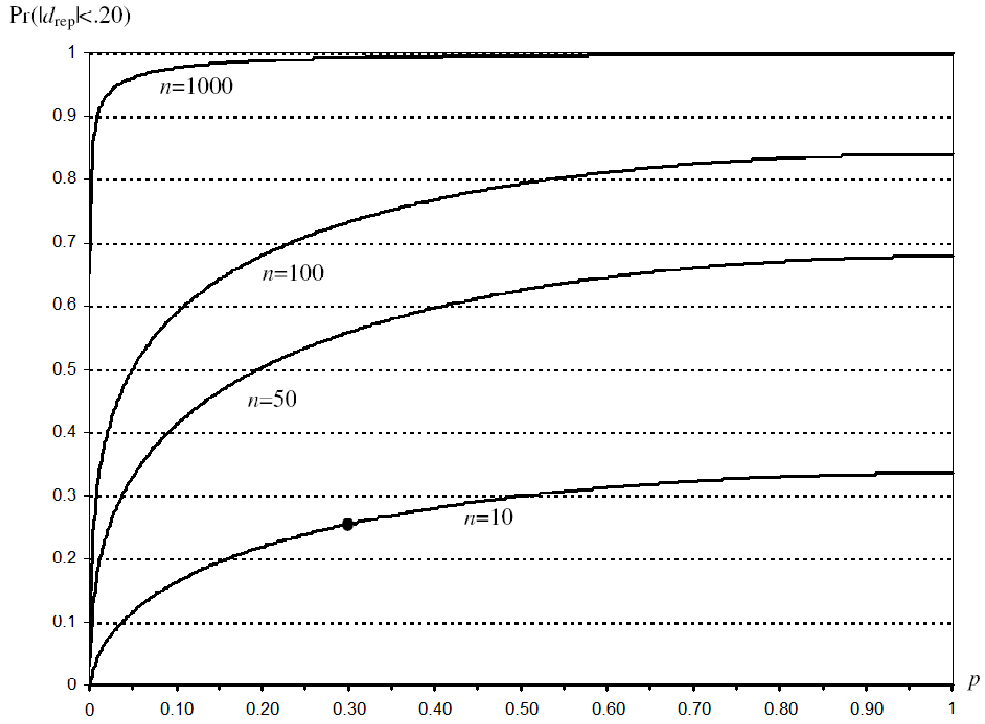

Figure 1. Examples of predictive probabilities of finding $|d_{rep}|<.20$ in a replication as a function of two-tailed *p*-value and number of paired differences (n = 10, 50, 100, 1000). The point corresponds to the CC example: $n = 10$, $p = .30$, $\Pr(|d_{rep}|<.20) = .255$ (The CC example is drawn from Conway and Christiansen, 2006).

What is the probability the upper confidence limit for μ/σ *in a replication does not exceed a given value?* Suppose that we are interested by the predictive probability that the upper limit of the 95% CI for μ/σ does not exceed 1.0. The 2.5% upper point of the noncentral *t* distribution with 9 *df* and noncentrality parameter $1.0\sqrt{10}$ is 1.162, hence the predictive probability given by

$$\Pr[K'_{(9,9)}(1.10/\sqrt{2}) < 1.162/\sqrt{2}] = .516. \qquad (14)$$

What is the probability the next replication mean will fall within the original 95% CI? Cumming, Williams and Fidler (2004) asked researchers to answer this question. Assuming σ known we find .834 (the probability that a Normal variable exceeds $1.96/\sqrt{2}$ in absolute value), which is precisely the value reported by the authors: "on average, just 5 out of 6 replication means (83.4%) will fall within an original 95% CI." (page 218). For σ unknown, the predictive probability is .856 (the probability that a Student *t* variable with 9 *df* exceeds $2.262/\sqrt{2}$ in absolute value: see Appendix B), which depends only on the number of *df*. This is the value given by Equation 8 (page 222) of Cumming and Maillardet (2006). However these authors gave a frequentist justification that involves the fact that the sampling distribution of $(x\text{-}x_{rep})/(s\sqrt{2/n})$ is a *t* distribution with ν *df*[10], hence a free parameter distribution. This distribution coincides with the predictive distribution in (8), but the frequentist interpretation involves repetitions

[10] This result is given in an equivalent form in Equation 7 of Cumming and Maillardet (2006).

including not only a sample of a replication, but also a sample of the experiment itself (x and s being also treated as random variables). The frequentist answer is numerically the same as the Bayesian predictive one. However, it must be realized that the frequentist and Bayesian procedures answer two different questions. The frequentist probability is the proportion of repeated pairs of samples such that the mean of the replication sample falls within the 95% CI computed from its associated experiment sample, *and not within the original CI*.

What is the probability the replication CI will contain the observed mean? This is the symmetrical question of the Cumming, Williams and Fidler question. As it can be expected for symmetry reasons the two probabilities are equal. Indeed, in this case $L = x_{obs}$, so that the predictive distributions for both the lower and upper limits are given by the central $K'_{(\nu,\nu)}(0)$ distribution, which is the t distribution with ν *df* (see B5 and B7).

What is the probability all the individual differences in the replication will be positive? In addition to the above "routine" procedures, simulation techniques allow to answer very specific questions. For instance, in the case of a within-subjects design, we may be interested in questions about individual differences. So, for the CC example, simulating 10^6 replicated samples, we get the predictive probability that the ten individual differences in the replication will be positive: .028. For instance, there is a probability .188 that at least 9 differences will be positive, a probability .819 that at least 5 differences will be positive. Etc.

Using Predictive Probabilities for Designing Experiments

The standard frequentist power approach for determining the sample size of a planned experiment relies on the choice of suitable values for the parameters (the effect size and the nuisance parameters), which implies more or less arbitrary choices. In contrast, the Bayesian approach is able to take into account the uncertainty about these parameters. The available information is expressed by a probability distribution, from which we compute in a very natural way the predictive probability that the future experiment will be in some sense conclusive. Some relevant references are: Brown, Herson, Atkinson and Rozell (1987); Lecoutre, Derzko and Grouin (1995); Lecoutre (2001, 2008); Spiegelhalter, Abrams and Myles (2004); Grouin, Coste, Bunouf and Lecoutre (2007).

How large does a future experiment need to be? In particular the available information can be the data of a preliminary experiment. If the objective of the planned experiment is to get a statistically significant result, the required predictive probability is analogous to p_{srep}, but is not restricted to a replication. By varying the future sample size, we get a predictive curve, expressing this probability as a function of the sample size. This curve can be used exactly as power curves for determining sample sizes that appear as limiting cases when the uncertainty about parameters vanishes. Of course, we may prefer a more interesting conclusion about effect sizes and consider predictive probabilities on interval estimates. For instance we can determine the minimal sample size needed to have a reasonable chance of obtaining a lower limit larger than a given relevant value. Furthermore, if available, the results of several earlier experiments can be combined to give the appropriate distribution about parameters.

Using Predictive Probabilities for Monitoring Experiments

Using realistic procedures to determine the sample size and searching for conclusions about effect sizes can considerably increase the cost of an experiment, making interim analyses particularly desirable. The predictive approach is a very appealing method (Baum, Houghton, & Abrams 1989) to aid the decision to stop an experiment at an interim stage. The Bayesian biostatistics literature may provide psychologists with useful methodological and technical tools for implementing predictive probabilities. Some relevant references are: Baum, Houghton and Abrams (1989); Spiegelhalter, Freedman and Parmar (1994); Lecoutre, Derzko and Grouin (1995); Johns and Andersen (1999); Lecoutre, Mabika and Derzko (2002); Berry (2005); Dmitrienko and Wang (2006); Lecoutre (2008).

When to stop? Given the available data at this stage, the predictive probabilities about the future planned data can be used to evaluate the chance that the final result for the whole data will be conclusive, or on the contrary inconclusive. On the one hand, if the predictive probability that it will be conclusive appears poor, it can be used as a rule to abandon the experiment. On the other hand, if the predictive probability is sufficiently high, this suggests we would stop the experiment early and conclude.

Note that interim analyses are a special kind of missing data imputation. More generally, predictive probabilities are also a valuable tool for missing data imputation.

Conclusion

The predictive idea is central in experimental investigations as the essence of science is replication. So Killeen's (2008, p. 120) objective "to validate observations through prediction and replication" appears as a desirable project. What are the pros and cons of this approach? On the negative side, we have identified some major problems concerning Killeen's p_{rep}. On the positive side, the Bayesian notion of predictive distribution is a fundamental tool for a better understanding of sample fluctuations. It provides a coherent statistical methodology for planning, monitoring and analyzing experiments.

On the Negative Side

In many articles p_{rep} appears to be incorrectly computed, due to the confusion between one-tailed and two-tailed p-values. No progress can be made without accepting the risk of computation error. Of course, this should not be used as a scientific argument to discard the use of a new statistical procedure.

Experimental findings strongly suggest that a common misconception is to confuse p_{rep}, the predictive probability of a same-sign result, with p_{srep}, the predictive probability of a significant result. Verbal explanations such as "be quite clear, however, that this [p_{rep} = .89] does not mean that the chance a replication will be statistically significant [p_{srep}] is .89" (Cumming, 2005, p. 1004) could have no more impact on such misunderstandings than the constant warnings about the misinterpretations of NHSTs. A related misconception is to interpret p_{rep} as "the probability that the experiment will give the same effect size."

Another objection is that predictive probabilities about replications engage us in an endless process: if p_{rep} or p_{srep} becomes a statistical output of an experiment, why not consider a predictive probability about these quantities (for instance the probability that p_{rep} will be larger

that .80 in a replication), and so on. This reminds us that predictive probabilities of replication are meaningful but that they cannot be considered as an end *per se*.

Last, a discouraging finding is that reporting p_{rep} seems to have little impact on the way the authors interpret their data. However, it must be emphasized that other changes that have been imposed in experimental publications, such as reporting confidence intervals, have also very little impact. For instance, Kotur (2006, p. 167) observed that "Indian Journal of Anaesthesia's editorial instruction to authors to report CIs and not p values in their studies, 3 years back has been largely ineffective: very few authors or no authors are referring to CIs when presenting their results." Indeed, most authors continue to focus on the statistical significance of the results, only wondering whether the CI includes the null hypothesis value, rather than on the full implications of confidence intervals. In other words, CIs are "simply used to do NHST" (Fidler, Thomason, Cumming, Finch & Leeman, 2004, p. 120).

On the Positive Side

Fiducial Bayesian predictive procedures provide different ways to summarize "what the data have to say". They include, among others, not only Killeen's p_{rep} for a replication, but also the traditional frequentist procedures when considering predictions about extremely large samples.

Correcting misconceptions. Predictive probabilities allow researchers to be aware of misconceptions about the replication of experiments. The confusion between p_{rep} and p_{srep} can be avoided by computing the two probabilities. This should also prevent the replicability fallacy, making explicit that 1-p is different from the predictive probability of a significant result in a replication. More generally, predictive probabilities can serve to understand how future replications vary and can correct misconceptions linked to the representativeness heuristic.

Analyzing and interpreting data. The predictive approach emphasizes the need to think hard about the information provided by the data in hand (what would happen if the experiment were to be replicated?) instead of applying ready-made procedures. In actual fact, the predictive probability of any event of interest about future data can be computed.

Moreover, predictive probabilities can be used to get inferences about parent parameters, with a more operational interpretation. This results from the fact that the parent effect size can be seen as the effect size that will be observed in an infinite, in practice extremely large, future sample. We get the fiducial Bayesian interpretation of the one-tailed p-value: 1-p is the predictive probability of a same-sign effect in a future *extremely large* data set (which is again the posterior probability of a same-sign effect in the population). This should prevent the misconception, called by Carver (1978) the *replicability fallacy*, which is to interpret 1-p as the predictive probability of a significant result in a replication.

This Bayesian interpretation makes clear that the p-value and p_{rep} only address questions about the sign of the effect. These questions are of limited interest and should be completed (or even replaced) with questions about effect sizes (standardized or not). Predictive probabilities give direct answers to these questions. An important feature is the interpretation of the usual confidence interval in terms of Bayesian probabilities: for instance in the CC example "there is a 95% probability of the future standardized difference in an extremely large sample (or again the population standardized difference) being included between the fixed bounds of the interval -.61 and +1.48" (conditionally on the data).

Defenders of frequentist methods should be reassured by the fact that p-values and confidence intervals can be used with the benefits of both the frequentist and fiducial Bayesian interpretations (in terms of posterior and of predictive probabilities) and without worrying about

the "correct justification." Furthermore, if pre-specified regions of the effect are of interest, their Bayesian probabilities can be obtained. For instance, for the CC example, let us suppose that we adopt the conventional Cohen's criteria about the smallness of a standardized difference. Then, for an extremely large sample size, the results can be summarized as follows: "there is a 28.7% predictive probability of a small standardized difference (less than .2 in absolute value), a 66.4% probability of a non small positive difference (more than +.2), and a 4.9% probability of a non small negative difference (less than -.2)." For a replication, these probabilities are respectively 22.5%, 62.3% and 12.2%. Such a statement has no frequentist counterpart.

Designing and monitoring experiments. We agree with Rozeboom (1960) that "the primary aim of a scientific experiment is not to precipitate decisions" (p. 420). However, we must also recognize that decisions are part of the scientific activity. In particular, a researcher must decide to replicate or not to replicate an experiment. In this perspective the probabilities of replication p_{rep} and p_{srep} can serve as routine procedures to help the decision. More sophisticated predictive procedures are available and give a very appealing method for more precise decisions such as to decrease or to increase the number of observations in the future experiment, to stop an experiment for futility, etc.

Concluding Remarks

Killeen's p_{rep} can be viewed as one of the many attempts to improve the habitual ways of analyzing and reporting experimental data. An increasingly widespread opinion is that effect size estimates and their confidence intervals should be reported in addition or in place of null hypothesis significance tests. The role of the planning of experiments (how many subjects?) is also stressed and power computations are often recommended. However, these attempts have not yet been really successful. Our analysis is that the frequentist approach is unable to provide a conceptually coherent statistical methodology. We argued that only the Bayesian approach can give researchers a real possibility of thinking sensibly about statistical inference problems and behaving in a more reasonable manner in the presentation and interpretation of results (Lecoutre, Lecoutre & Poitevineau, 2001; Lecoutre, 2006b). Since most people use "inverse probability" statements to interpret NHST and confidence intervals, probabilistic concepts involved in the Bayesian approach, in particular the Bayesian definition of probability, are already – at least implicitly – familiar to researchers.

Bayesian predictive probabilities, because they relate observables between each other, are very intuitive and even more natural than posterior probabilities about parameters (Bernard, 2000). They should be an important part of the statistical teaching and training of psychologists. We do not claim that they should replace a genuine inference about population parameters, such as an interval estimate for an appropriate effect size measure. However, they may wonderfully complement it. As emphasized by Gigerenzer (1998), "we need statistical thinking, not rituals". A researcher cannot be unconcerned about "what would happen if additional subjects were to be included into the experiment?", "what would be the conclusion for the data of these future subjects?", "what would be the conclusion for the whole data?", or "what would happen if this experiment were to be repeated?" Asking and answering such questions goes beyond the ritualized statistical procedures, and is likely to influence the way the authors of scientific papers interpret experimental findings and conduct their experiments. Predictive probabilities are an unavoidable part of statistical thinking and the time is come to take them seriously.

References

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Vaughn Becker, D., Griskevicius, V., Maner, J. K., and Schaller, M. (2006). They all look the same to me (unless they're angry): from out-group homogeneity to out-group heterogeneity. *Psychological Science*, *17*, 836- 840.

Baum, M., Houghton, J., and Abrams, K. R. (1989). Early stopping rules: clinical perspectives and ethical considerations. *Statistics in Medicine*, *13*, 1459-1469.

Bayarri, M. J., and Berger, J. P. (1997). Measures of surprise in Bayesian analysis. ISDS Discussion Paper 97-46, Duke University. Retrieved December 1, 2008, from http://citeseer.ist.psu.edu/459777.html.

Bayarri, M. J., and Berger, J. O. (2004). the interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*, 58-80.

Berger, J. (2004). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*, 1-17.

Bernard, J.-M. (2000). Bayesian inference for categorized data. 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. 159-226). Bern, Switzerland: Peter Lang.

Berry, D. (1991). Experimental design for drug development: A Bayesian approach. *Journal of Biopharmaceutical Statistics*, *1*, 81-101.

Berry, D. (2005). Introduction to Bayesian methods III: Use and interpretation of Bayesian tools in design and analysis. *Clinical Trials*, *2*, 295-300.

Box, G. E .P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison Wesley.

Brown, B. W., Herson J., Atkinson E. N., and Rozell M. E. (1987). Projection from previous studies - A Bayesian and frequentist compromise. *Controlled Clinical Trials*, *8*, 29-44.

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

Conway, C. M., and Christiansen, M. H., (2006). Statistical learning within and between modalities. *Psychological Science*, *17*, 905-912.

Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, *16*, 1002-1004.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*, 286-300.

Cumming, G., and 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.

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

Cumming, G., Williams, J., and Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*, 299–311.

Dixon, J., Durrheim, K., and Tredoux, C. (2007). Intergroup contact and attitudes toward the principle and practice of racial equality. *Psychological Science*, *18*, 867-872.Dmitrienko, A., and Wang, M. D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine*, *25*, 2178-2195.

Fidler, F., Thomason, N., Cumming, G., Finch, S., and Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. *Psychological Science*, *15*, 119-26.

Fisher, R. A. (1990/1970). *Statistical Methods for Research Workers* [14th ed. 1970, re-issue edited by J. H. Bennet with a foreword by F. Yates, *Statistical Methods, Experimental Design and Scientific Inference*]. Oxford: Oxford University Press.

Fisher, R. A. (1990/1973). *Statistical Methods and Scientific Inference* [3rd ed.1973, re-issue edited by J. H. Bennet with a foreword by F. Yates, *Statistical Methods, Experimental Design and Scientific Inference*]. Oxford: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis* (2nd ed.). London: Chapman and Hall.

Gerits, N., Van Belle, W., and Moens, U. (2007). Transgenic mice expressing constitutive active MAPKAPK5 display gender-dependent differences in exploration and activity. *Behavioral and Brain Functions*, 2007, *3*:58.Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199-200.

Goodman, S. N. (2002). Author's reply. *Statistics in Medicine*, *21*, 2445-2447.

Grouin J.-M., Coste M., Bunouf P., and Lecoutre B. (2007). Bayesian sample size determination in non-sequential clinical trials: Statistical aspects and some regulatory considerations. *Statistics in Medicine*, *26*, 4914-4924.

Guttman, L. (1977). What is not what in statistics? *The Statistician*, *26*, 81-107.

Johns, D. and Andersen, J. S. (1999). Use of predictive probabilities in phase II and phase III clinical trials. *Journal of Biopharmaceutical Statistics*, *9*, 67-79.

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

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, England: Cambridge University Press.

Kahneman, D., and Tversky, A. (1972). Subjective probability: A judgement of representativeness. *Cognitive Psychology*, *3*, 430-454.

Killeen, P. R. (2005a). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*, 345-353.

Killeen, P. R. (2005b). Replicability, confidence, and priors. *Psychological Science*, *16*, 1009-1012.

Killeen, P. R. (2008). Replication statistics as a replacement for significance testing: Best practices in scientific decision-making. In J. W. Osborne (Ed.), *Best practices in quantitative methods*. Thousand Oaks, CA: Sage Publishing (pp. 103-124).

Kotur, P. F. (2006). Statistics In Biomedical Journals: 'Science' has the right to be written well! Editorial. *Indian Journal of Anaesthesia*, *50*, 166-168.

Lecoutre, B. (1984). *L'Analyse Bayésienne des Comparaisons* [*The Bayesian Analysis of Comparisons*]. Lille, France: Presses Universitaires de Lille.

Lecoutre, B. (1985). Reconsideration of the F test of the analysis of variance: The semi-Bayesian significance tests. *Communications in Statistics: Theory and Methods*, *14*, 2437-2446.

Lecoutre, B. (1999). Two useful distributions for Bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference*, *79*, 93-105.

Lecoutre, B. (2000). From significance tests to fiducial Bayesian inference. 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. 123-157). Bern, Switzerland: Peter Lang.

Lecoutre, B. (2001). Bayesian predictive procedure 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. (2006a). Training students and researchers in Bayesian methods. *Journal of Data Science*, *4*, 207-232.

Lecoutre B. (2006b). And if you were a Bayesian without knowing it? *In* A. Mohammad-Djafari (Ed.), *26th workshop on Bayesian inference and maximum entropy methods in science and engineering* (pp. 15-22). Melville: AIP Conference Proceedings Vol. 872.

Lecoutre, B. (2007). Another look at confidence intervals for the noncentral t distribution. *Journal of Modern Applied Statistical Methods*, *6*, 155-164.

Lecoutre, B. (2008). Bayesian methods for experimental data analysis. In C. R. Rao, J. Miller & D. C. Rao (Eds.), *Handbook of statistics: Epidemiology and medical statistics* (Vol 27) (pp. 775-812). Amsterdam, Nederland: Elsevier.

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

Lecoutre, B., Lecoutre, M.-P., and 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., Mabika, B., and Derzko, G. (2002). Assessment and monitoring in clinical trials when survival curves have distinct shapes in two groups: a Bayesian approach with Weibull modeling illustrated. *Statistics in Medicine*, *21*, 663-674.

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

Lecoutre, M.-P., Poitevineau, J., and Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology*, *38*, 37-45.

Macdonald, R. R (2005). Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen (2005). *Psychological Science*, *16*, 1007-1008.

Meng, X.-L. (1994). Posterior predictive *p*-values. *The Annals of Statistics*, *22*, 1142-1160.

Murphy, M. C., Steele, C. M., and Gross, J. J. (2007). Signaling Threat: How Situational Cues Affect Women in Math, Science, and Engineering Settings. *Psychological Science*, *18*, 879-885.

M. *Part 1, Some Experimental Findings About the Interpretation*

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. New York: Wiley.

Poitevineau J., and 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.

Poitevineau J., and Lecoutre B. (2008). Implementing Bayesian predictive procedures: The K-prime and K-square distributions. *Computational Statistics & Data Analysis*, in press.

Rosenthal, R., and Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, *5*, 329-334.

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

Rouanet H., and Lecoutre B. (1983). Specific inference in ANOVA: From significance tests to Bayesian procedures. *British Journal of Mathematical and Statistical Psychology*, *36*, 252-268.

Rozeboom, W. W (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, *57*, 416-428.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. New York: Wiley 2004.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society A*, *157*, 357-416.

Student (1908). The probable error of a mean. *Biometrika*, *6*, 1-25.

Tversky, A., and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 237-251.

Zabell, S. L. (1992). R. A. Fisher and the fiducial argument. *Statistical Science*, *7*, 369-387.

Appendix A: The *K-Prime* Distribution

The *K*-prime distribution was studied in detail in Lecoutre (1999). It includes all the other involved distributions as particular cases[11]:

$K'_{(\nu1,\nu2)}(0) \equiv t_{(\nu2)}$ (usual *t*)

$K'_{(\infty,\nu2)}(a) \equiv t'_{(\nu2)}(a)$ (noncentral *t*)

$K'_{(\nu1,\infty)}(a) \equiv \Lambda'_{(\nu1)}(a)$ (lambda-prime)

$K'_{(\infty,\infty)}(a) \equiv N(a,1)$ (normal)

An algorithm for computing its cumulative distribution function was given in Poitevineau and Lecoutre (2008) and a computer program is freely available (see Section 3).

Appendix B: Predictions for an Unstandardized Contrast

The sampling distribution of the unstandardized contrast x_{rep} in a future sample is a normal distribution. It follows that the resulting averaged predictive distribution (given x_{obs} and s_{obs}) is a Student *t* distribution. For the CC example, this distribution is centered on x_{obs} and has a scale factor $s_{obs}\sqrt{2/n}$. For large values of ν (or for σ known), it is a normal distribution with mean x_{obs} and standard deviation $\sigma\sqrt{2/n}$. It follows that the predictive distribution of x_{rep} given x_{obs} and t_{obs} is such that:

$$(x_{rep}\text{-}x_{obs})/(s_{obs}\sqrt{2/n}) \sim t_{(\nu)} \qquad \text{(B1)}$$

A general form for an unstandardized contrast (assuming $x_{obs} \neq 0$) is:

$$(x_{rep}\text{-}x_{obs})/(x_{obs}\text{-}t_{obs}) \sim t_{(\nu)} \qquad \text{(B2)}$$

[11] Note again that exact tests and confidence limits for the correlation coefficient ρ can be computed from the *K*-prime distribution (Poitevineau & Lecoutre, 2008).

The predictive probability that x_{rep} exceeds C is:

$$\Pr(x_{rep} > C) = \Pr[t_{(\nu)} < ((x_{obs}-C)/x_{obs})t_{obs}/\sqrt{2}\,], \qquad \text{(B3)}$$

which is a generalization of (4) and is of course p_{rep} in the particular case C=0.

What is the probability the next replication mean will fall within the observed 95% CI? The 100(1-α)% observed CI for μ is $x_{obs} \pm (s_{obs}/\sqrt{n})t_{c(\nu,\alpha)}$, hence for the CC example [-5.29,+15.29], computed from $x_{obs} = 5.0$ and $s_{obs} = 14.38$. The next replication will fall within this CI if $|x_{rep}-x_{obs}|/(s_{obs}\sqrt{2/n}) < t_{c(\nu,\alpha)}\sqrt{2}$. The corresponding predictive probability, derived from (B1) is $\Pr(|t_{(9)}| > 2.262/\sqrt{2}) = .856$ ($t_{c(9,.05)} = 2.262$).

Prediction Interval for an Unstandardized Contrast

It follows from (B1) that a 100(1-α)% prediction interval for the mean difference x_{rep} is:

$$x_{obs} \pm (s_{obs}\sqrt{2/n})t_{c(\nu,\alpha)} \qquad \text{(B4)}$$

Note the similarity with the usual 100(1-α)% CI for μ: as a general result for an unstandardized contrast between means, the limits are simply multiplied by $\sqrt{2}$. For the CC example, from the 95% CI [-5.29,+15.29] for μ, we get the 95% prediction interval for x_{rep}: [$-5.29\sqrt{2} = -9.55$, $15.29\sqrt{2} = +19.55$].

Predictions About the Confidence Limits of an Unstandardized Contrast

The lower limit of the 100(1-α)% CI for μ in a replication is $x_{rep} - (s_{rep}/\sqrt{n})t_{c(\nu,\alpha)}$. The predictive probability that this limit exceeds a given value L is equal to the probability that $(x_{rep}-L)/(s_{rep}/\sqrt{n})$, i.e. the t statistic for testing the null hypothesis $H_0 : \mu = L$, exceeds $t_{c(\nu,\alpha)}$. It is given by

$$\Pr\Big[K'_{(\nu,\nu)}\big[[(x_{obs}-L)/(s_{obs}/\sqrt{n})]/\sqrt{2}\,\big] > t_{c(\nu,\alpha)}/\sqrt{2}\,\Big]. \qquad \text{(B5)}$$

hence the general formula for a contrast between means (assuming $x_{obs} \neq 0$):

$$\Pr\big[K'_{(\nu,\nu)}[(1-L/x_{obs})t_{obs}/\sqrt{2}\,] > t_{c(\nu,\alpha)}/\sqrt{2}\,\big], \qquad \text{(B6)}$$

which is a straightforward generalization of (5).

For $L = 0$, it is the probability of a same-sign significant replication $p_{srep}(\alpha)$ if $x_{obs} > 0$, and the probability of an opposite-sign significant replication if $x_{obs} < 0$. More generally, it is the predictive probability of an unstandardized effect larger than L and such that the null hypothesis $\mu = L$ is rejected at two-tailed level α.

In the same way, the predictive probability that the upper limit $x_{rep} + (s_{rep}/\sqrt{n})t_{\nu(\alpha)}$ does not exceed U is given by

$$\Pr\big[K'_{(\nu,\nu)}[(1-U/x_{obs})t_{obs}/\sqrt{2}\,] < -t_{c(\nu,\alpha)}/\sqrt{2}\,\big], \qquad \text{(B7)}$$

It is the predictive probability of an unstandardized effect smaller than U and such that the null hypothesis $H_0 : \mu = U$ is rejected at two-tailed level α.

What is the probability the lower confidence limit for μ *in a replication exceeds a given value?* The predictive probability of a positive lower limit for the 95% CI in a replication is $p_{srep}(.05) = .230$. Note that, by symmetry, we find that $p_{srep}(.05)$ is also the predictive probability that the upper limit does not exceed $2\times5.0 = 10.0$, the Rosenthal and Rubin (1994) counternull value of the effect size. Suppose that we are interested by the predictive probability that the lower limit for μ does not exceed a given value in the negative direction. For instance, the

probability that the limit will be greater than -5.0 (the opposite of the observed difference) is given by (13): $\Pr(K'_{(9,9)}(2\times 1.10/\sqrt{2})) > 2.262/\sqrt{2}) = .485.$