independent, the applications were restricted to situations where the data $x = (x_1, \ldots, x_n)$ were split according to $s(x) = (x_1, \ldots, x_{n_*})$ and $r(x) = (x_{n_*+1}, \ldots, x_n)$. The approach taken in this paper will likely produce better choices of (r, s) for model checking in a number of situations but the use of the splits used in Evans (1997), perhaps with random splitting, will likely be the only feasible ones in many contexts.

The use of the tail probabilities $P(T(X) \geq t_{obs})$, for some probability measure P, does not work well in capturing the idea of a surprising observation having occurred when this probability is small. This is because t_{obs} could be an extreme value in the left tail or, in general be in a low probability region such as near an anti-mode and $P(T(X) \geq t_{obs})$ will not indicate this. For this reason authors such as Weaver, Good and Box have worked instead with the density of T as this corrects for this problem. Using the density, however, destroys the invariance of the measure. The observed relative surprise, introduced in Evans (1997) corrects for both of these problems.

DENNIS V. LINDLEY *(Minehead, UK)*

Objections are often raised to the Bayesian approach because of its dependence on the prior. It is not so often recognized that the p-value can equally be criticized because of its dependence on the sample space. One can produce, for a given data set, a range of p-values by varying the sample space. It follows that since, in most practical cases, the sample space for a given data set is ill-defined, the p-value is also ambiguous. In particular, it can only be considered as a measure of surprise when the sample space is unambiguous. My personal view is that p-values should be relegated to the scrap heap and not considered by those who wish to think and act coherently.

XIAO-LI MENG *(The University of Chicago, USA)* and
ANDREW GELMAN *(Columbia University, USA)*

A paper containing multiple ideas is always fun to read. The main idea of Section 2, namely converting a p-value into a lower bound for Bayes factors is quite intriguing, especially considering that Bayes factors and p-value type measures answer two different statistical questions – a model can have a high Bayes factor compared to its stated competitors but still poorly fit important aspects of observed data. What's unclear to us, in general, is *which* p-value can be used to construct a useful lower bound given that a p-value is a functional of test statistics (or more generally *discrepancies*, i.e., $T(X, \theta)$), choices of replications (see below), etc. Perhaps the p-value from the likelihood ratio test (or the conditional likelihood ratio as defined in Meng (1994))?

Regarding the central theme of Section 3, we view p-value as *a* measure of *discrepancy* between the posited model and the data being analyzed, as we emphasized strongly in Meng (1994) and Gelman, Meng, and Stern (1996). While it might be a semantic matter to some, we prefer the term *discrepancy* because it honestly reflects what a small p-value tells us: the data and the model do not seem to see eye to eye in a specified way, but we cannot tell you which to blame! The phrase "surprise in the data" seems to carry the impression that the problem is with the data (e.g., an "unlucky" sample) and the standard practice with hypothesis testing always emphasizes the rejection of the posited hypothesis, not the data. While it is obviously desirable to pinpoint the sources of the discrepancy, p-value type measures simply cannot tell us whether the problem is with the data, or the model, or, as is more likely, both! If one's goal is simply to fit the data, then of course the source is always the model. But with scientific inferences, the problem can be far more complicated – Example 2 is a good illustration.

Viewed as discrepancy measure, Example 4 can be readily understood. A large value of $|\bar{x}_{obs}|$ does not necessarily indicate a large discrepancy between the data and the posited model

$N(0, \sigma^2)$ unless we know for sure the value of σ^2, which is precisely why the lower bound on the posterior predictive p-value given in (4.11) monotonically decreases to zero as n approaches infinity. This also suggests that a relative measure such as $|\bar{x}|/s$ would be more useful for detecting the discrepancy in the mean. Indeed, with this choice of discrepancy the posterior predictive p-value would be identical to that from the classical (two-sided) t test, i.e., the p-value given in (4.9).

Incidentally, (4.9) can be obtained under the posterior predictive framework even when using $|\bar{x}|$ if we ignore the null value $\mu = 0$ when computing the posterior for σ^2 under the same model as used in the paper. Note that although Example 4 states that the null is $N(0, \sigma^2)$, any model checking procedure based solely on $\bar{x}$ and s cannot check the normality assumption – indeed, the classical t test and other methods discussed in Example 4 are robust to the normality assumption (unless n is very small). So it would be better to cast this problem as checking the mean parameter $\mu = 0$ verses $\mu \neq 0$, which was the original formulation given in Meng (1994). With this formulation, the classical answer is obtained if we use the marginal posterior $p(\sigma^2|x_{obs})$ instead of the the conditional one $p(\sigma^2|x_{obs}, \mu = 0)$; see Meng (1994) for details. The problem with using $p(\sigma^2|x_{obs}, \mu = 0)$, from the point of view of testing $\mu = 0$ (not of checking discrepancy in $|\bar{x}_{obs} - 0|$), is that it can grossly overestimate σ^2 when $\mu \neq 0$ and thus leads to the very conservative nature of the p-value given in (4.11).

This example makes it clear that the choice of test/discrepancy is important and is confounded with the choice of replication. Throughout the literature of posterior predictive checking (e.g., Rubin, 1984; Meng, 1994; Gelman, Meng and Stern, 1996), these points are always emphasized. For example, in Gelman, Meng and Stern (1996), we explicitly define $f(x|\theta, A)$ with A being an auxiliary statistics, as the authors mentioned. What is proposed in Section 4 is to condition on such an A (or in authors' notation, U) instead of the full data when finding the posterior of θ. Such conditioning is in the same spirit as conditioning on the classical ancillary statistics (for the parameter fixed by the null model, i.e., μ, not σ^2 in Example 4), which is in the right direction for a frequentist in the mind of a Bayesian because it is towards full conditioning, but is in the opposite direction when one is already doing Bayesian full conditioning. It would be better to resolve the "power" issue through the choices of discrepancy and the sampling replication $f(x|\theta, A)$. Even the marginal (e.g., $p(\sigma^2|x_{obs})$) verses conditional (e.g., $p(\sigma^2|x_{obs}, \mu = 0)$) approach is unsatisfactory because once we allow ourselves to not fully condition on the null model when computing the p value, we would need a principle to decide to what extent the null should be conditioned upon.

Of course, mathematically speaking, having more flexibility implies possibly better optimality in terms of frequentist operating characteristics of the resulting procedures. We look forward to seeing more convincing examples of the utility of the conditional predictive approach (incidentally we think the term *partially conditional predictive* would be more precise than *conditional predictive* because the posterior predictive approach is conditional, in fact, full conditional predictive approach with the authors' use of the word "conditional"). Example 4 would be theoretically more revealing if the "perfectly satisfactory" answer (i.e., the classical t test) could only be obtained under the proposed partial conditioning approach – with the current example, the answer can be obtained via any predictive approach, from no conditioning (i.e., prior predictive) to full conditioning (i.e., posterior predictive), since $|x|/s$ is pivotal.

MICHEL MOUCHART *(Université Catholique de Louvain, Belgium)*

(i) That a quantification of "surprise" depends on the prior specification might be desirable for several reasons among which one should mention : a) in a Bayesian model, i.e. a joint probability on the observations and the parameters, the "sampling" component is as subjective, and liable to "doubts", as the "prior" components, b) as argued in p.316 (remark (iii)) in Florens