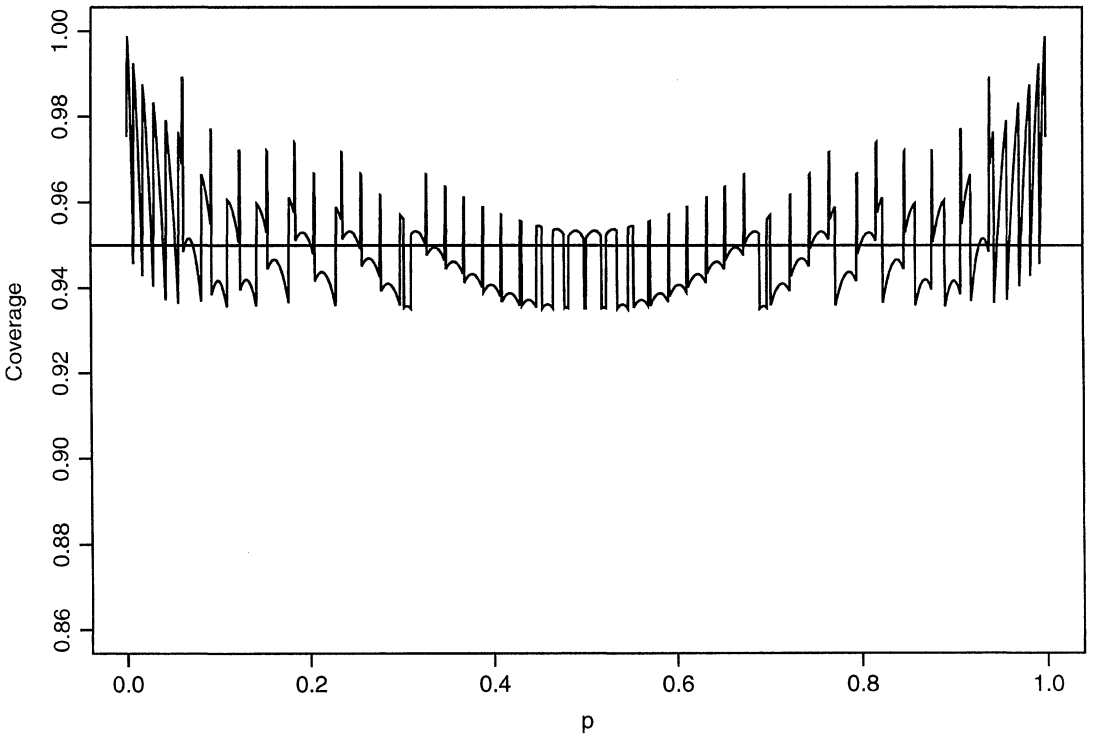

FIG. 2. *Coverage of nominal* 93.5% *symmetric Duffy–Santner* p *intervals for* $n = 50$.

Most of these would be extremely difficult to handle by the more brute force method of inverting acceptance sets. The first of these is the problem of computing simultaneous confidence intervals for $p_0 - p_i, 1 \leq i \leq T$ that arises in comparing a control binomial distribution with T treatment ones. The second concerns forming simultaneous confidence intervals for $p_i - p_j$, the cell probabilities of a multinomial distribution. In particular, the equal-tailed Jeffrey prior approach recommended by the author has strong appeal for both of these problems.

Finally, I note that the Wilson intervals seem to have received some recommendation as the method of choice in other elementary texts. In his introductory texts, Larson (1974) introduces the Wilson interval as the method of choice although he makes the vague, and indeed false, statement, as BCD show, that the user can use the Wald interval if "n is large enough." One reviewer of Santner (1998), an article that showed the coverage virtues of the Wilson interval compared with Wald-like intervals advocated by another author in the magazine *Teaching Statistics* (written for high school teachers) commented that the Wilson method was the "standard" method taught in the U.K.

Rejoinder

Lawrence D. Brown, T. Tony Cai and Anirban DasGupta

We deeply appreciate the many thoughtful and constructive remarks and suggestions made by the discussants of this paper. The discussion suggests that we were able to make a convincing case that the often-used Wald interval is far more problematic than previously believed. We are happy to see a consensus that the Wald interval deserves to be discarded, as we have recommended. It is not surprising to us to see disagreement over the specific alternative(s) to be recommended in place of

this interval. We hope the continuing debate will add to a greater understanding of the problem, and we welcome the chance to contribute to this debate.

A. It seems that the primary source of disagreement is based on differences in interpretation of the coverage goals for confidence intervals. We will begin by presenting our point of view on this fundamental issue.
We will then turn to a number of other issues, as summarized in the following list:

B. Simplicity is important.
C. Expected length is also important.
D. Santner's proposal.
E. Should a continuity correction be used?
F. The Wald interval also performs poorly in other problems.
G. The two-sample binomial problem.
H. Probability-matching procedures.
I. Results from asymptotic theory.

A. Professors Casella, Corcoran and Mehta come out in favor of making coverage errors always fall only on the conservative side. This is a traditional point of view. However, we have taken a different perspective in our treatment. It seems more consistent with contemporary statistical practice to expect that a $\gamma\%$ confidence interval should cover the true value *approximately* $\gamma\%$ of the time. The approximation should be built on sound, relevant statistical calculations, and it should be as accurate as the situation allows.

We note in this regard that most statistical models are only felt to be approximately valid as representations of the true situation. Hence the resulting coverage properties from those models are at best only approximately accurate. Furthermore, a broad range of modern procedures is supported only by asymptotic or Monte-Carlo calculations, and so again coverage can at best only be approximately the nominal value. As statisticians we do the best within these constraints to produce procedures whose coverage comes close to the nominal value. In these contexts when we claim $\gamma\%$ coverage we clearly intend to convey that the coverage is close to $\gamma\%$, rather than to guarantee it is at least $\gamma\%$.

We grant that the binomial model has a somewhat special character relative to this general discussion. There are practical contexts where one can feel confident this model holds with very high precision. Furthermore, asymptotics are not *required* in order to construct practical procedures or evaluate their properties, although asymptotic calculations can be useful in both regards. But the discreteness of the problem introduces a related barrier to the construction of satisfactory procedures. This forces one to again decide whether $\gamma\%$ should mean "approximately $\gamma\%$," as it does in most other contemporary applications, or "at least $\gamma\%$" as can be obtained with the Blyth–Still procedure or the Cloppe–Pearson procedure. An obvious price of the latter approach is in its decreased precision, as measured by the increased expected length of the intervals.

B. All the discussants agree that elementary motivation and simplicity of computation are important attributes in the classroom context. We of course agree. If these considerations are paramount then the Agresti–Coull procedure is ideal. If the need for simplicity can be relaxed even a little, then we prefer the Wilson procedure: it is only slightly harder to compute, its coverage is clearly closer to the nominal value across a wider range of values of p, and it can be easier to motivate since its derivation is totally consistent with Neyman–Pearson theory. Other procedures such as Jeffreys or the mid-P Clopper–Pearson interval become plausible competitors whenever computer software can be substituted for the possibility of hand derivation and computation.

Corcoran and Mehta take a rather extreme position when they write, "Simplicity and ease of computation have *no role* to play in statistical practice [italics ours]." We agree that the ability to perform computations by hand should be of little, if any, relevance in practice. But conceptual simplicity, parsimony and consistency with general theory remain important secondary conditions to choose among procedures with acceptable coverage and precision.

These considerations will reappear in our discussion of Santner's Blyth–Still proposal. They also leave us feeling somewhat ambivalent about the boundary-modified procedures we have presented in our Section 4.1. Agresti and Coull correctly imply that other boundary corrections could have been tried and that our choice is thus somewhat ad hoc. (The correction to Wilson can perhaps be defended on the principle of substituting a Poisson approximation for a Gaussian one where the former is clearly more accurate; but we see no such fundamental motivation for our correction to the Jeffreys interval.)

C. Several discussants commented on the precision of various proposals in terms of expected length of the resulting intervals. We strongly concur that precision is the important balancing criterion vis-á-vis coverage. We wish only to note that there exist other measures of precision than interval expected length. In particular, one may investigate the probability of covering wrong values. In a

charming identity worth noting, Pratt (1961) shows the connection of this approach to that of expected length. Calculations on coverage of wrong values of p in the binomial case will be presented in DasGupta (2001). This article also discusses a number of additional issues and presents further analytical calculations, including a Pearson tilting similar to the chi-square tilts advised in Hall (1983).

Corcoran and Mehta's Figure 2 compares average length of three of our proposals with Blyth–Still and with their likelihood ratio procedure. We note first that their LB procedure is not the same as ours. Theirs is based on numerically computed exact percentiles of the fixed sample likelihood ratio statistic. We suspect this is roughly equivalent to adjustment of the chi-squared percentile by a Bartlett correction. Ours is based on the traditional asymptotic chi-squared formula for the distribution of the likelihood ratio statistic. Consequently, their procedure has conservative coverage, whereas ours has coverage fluctuating around the nominal value. They assert that the difference in expected length is "negligible." How much difference qualifies as negligible is an arguable, subjective evaluation. But we note that in their plot their intervals can be on average about 8% or 10% longer than Jeffreys or Wilson intervals, respectively. This seems to us a nonnegligible difference. Actually, we suspect their preference for their LR and BSC intervals rests primarily on their overriding preference for conservativity in coverage whereas, as we have discussed above, our intervals are designed to attain approximately the desired nominal value.

D. Santner proposes an interesting variant of the original Blyth–Still proposal. As we understand it, he suggests producing nominal $\gamma\%$ intervals by constructing the $\gamma^*\%$ Blyth–Still intervals, with $\gamma^*\%$ chosen so that the average coverage of the resulting intervals is approximately the nominal value, $\gamma\%$. The coverage plot for this procedure compares well with that for Wilson or Jeffreys in our Figure 5. Perhaps the expected interval length for this procedure also compares well, although Santner does not say so. However, we still do not favor his proposal. It is conceptually more complicated and requires a specially designed computer program, particularly if one wishes to compute $\gamma^*\%$ with any degree of accuracy. It thus fails with respect to the criterion of scientific parsimony in relation to other proposals that appear to have at least competitive performance characteristics.

E. Casella suggests the possibility of performing a continuity correction on the score statistic prior to constructing a confidence interval. We do not agree with this proposal from any perspective. These "continuity-corrected Wilson" intervals have extremely conservative coverage properties, though they may not in principle be guaranteed to be everywhere conservative. But even if one's goal, unlike ours, is to produce conservative intervals, these intervals will be very inefficient at their normal level relative to Blyth–Still or even Clopper–Pearson. In Figure 1 below, we plot the coverage of the Wilson interval with and without a continuity correction for $n = 25$ and $\alpha = 0.05$, and the corresponding expected lengths. It is seems clear that the loss in precision more than neutralizes the improvements in coverage and that the nominal coverage of 95% is misleading from any perspective.

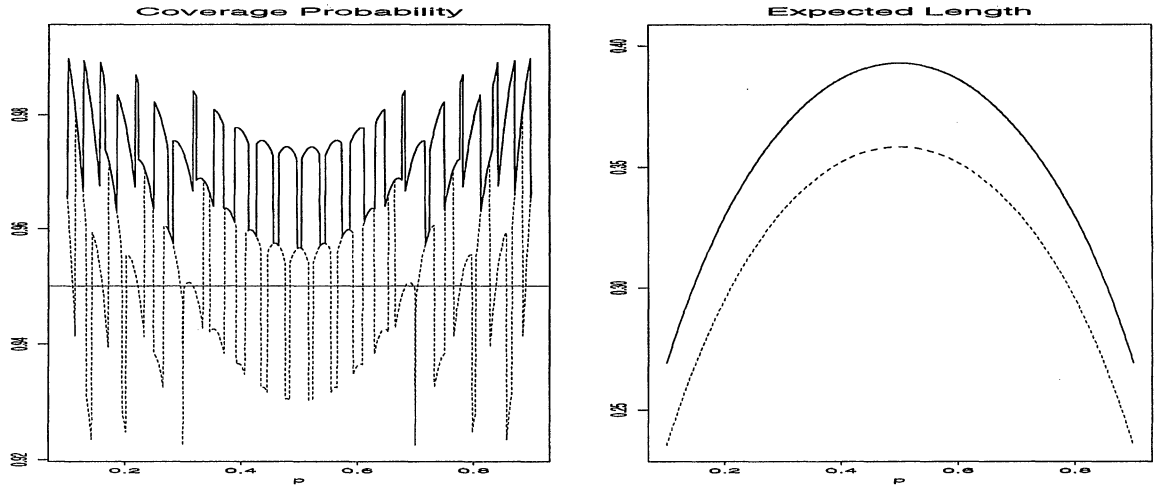

FIG. 1. *Comparison of the coverage probabilities and expected lengths of the Wilson (dotted) and continuity-corrected Wilson (solid) intervals for $n = 25$ and $\alpha = 0.05$.*

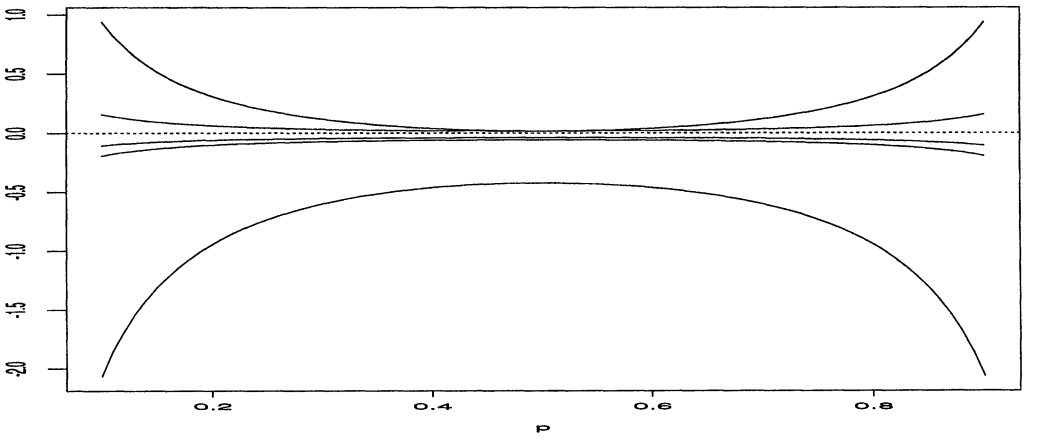

FIG. 2. *Comparison of the systematic coverage biases. The y-axis is* $nS_n(p)$. *From top to bottom: the systematic coverage biases of the Agresti–Coull, Wilson, Jeffreys, likelihood ratio and Wald intervals, with* $n = 50$ *and* $\alpha = 0.05$.

F. Agresti and Coull ask if the dismal performance of the Wald interval manifests itself in other problems, including nordiscrete cases. Indeed it does. In other lattice cases such as the Poisson and negative binomial, both the considerable negative coverage bias and inefficiency in length persist. These features also show up in some continuous exponential family cases. See Brown, Cai and DasGupta (2000b) for details.

In the three important discrete cases, the binomial, Poisson and negative binomial, there is in fact some conformity in regard to which methods work well in general. Both the likelihood ratio interval (using the asymptotic chi-squared limits) and the equal-tailed Jeffreys interval perform admirably in all of these problems with regard to coverage and expected length. Perhaps there is an underlying theoretical reason for the parallel behavior of these two intervals constructed from very different foundational principles, and this seems worth further study.

G. Some discussants very logically inquire about the situation in the two-sample binomial situation. Curiously, in a way, the Wald interval in the two-sample case for the difference of proportions is less problematic than in the one-sample case. It can nevertheless be somewhat improved. Agresti and Caffo (2000) present a proposal for this problem, and Brown and Li (2001) discuss some others.

H. The discussion by Ghosh raises several interesting issues. The definition of "first-order probability matching" extends in the obvious way to any set of upper confidence limits; not just those corresponding to Bayesian intervals. There is also an obvious extension to lower confidence limits. This probability matching is a one-sided criterion. Thus a family of two-sided intervals $[L_n, U_n]$ will be first-order probability matching if

$$Pr_p(p \le L_n) = \alpha/2 + o(n^{-1/2}) = Pr_p(p \ge U_n).$$

As Ghosh notes, this definition cannot usefully be literally applied to the binomial problem here, because the asymptotic expansions always have a discrete oscillation term that is $O(n^{-1/2})$. However, one can correct the definition.

One way to do so involves writing asymptotic expressions for the probabilities of interest that can be divided into a "smooth" part, S, and an "oscillating" part, Osc, that averages to $O(n^{-3/2})$ with respect to any smooth density supported within (0, 1). Readers could consult BCD (2000a) for more details about such expansions. Thus, in much generality one could write

$$(1) \quad \begin{aligned} &Pr_p(p \le L_n) \\ &\quad = \alpha/2 + S_{L_n}(p) + Osc_{L_n}(p) + O(n^{-1}), \end{aligned}$$

where $S_{L_n}(p) = O(n^{-1/2})$, and $Osc_{L_n}(p)$ has the property informally described above. We would then say that the procedure is first-order probability matching if $S_{L_n}(p) = o(n^{-1/2})$, with an analogous expression for the upper limit, U_n.

In this sense the equal-tail Jeffreys procedure is probability matching. We believe that the mid-P Clopper–Pearson intervals also have this asymptotic property. But several of the other proposals, including the Wald, the Wilson and the likelihood ratio intervals are not first-order probability matching. See Cai (2001) for exact and asymptotic calculations on one-sided confidence intervals and hypothesis testing in the discrete distributions.

The failure of this one-sided, first-order property, however, has no obvious bearing on the coverage properties of the two-sided procedures considered in the paper. That is because, for any of our procedures,

$$S_{L_n}(p) + S_{U_n}(p) = 0 + O(n^{-1}), \tag{2}$$

even when the individual terms on the left are only $O(n^{-1/2})$. All the procedures thus make compensating one-sided errors, to $O(n^{-1})$, even when they are not accurate to this degree as one-sided procedures.

This situation raises the question as to whether it is desirable to add as a secondary criterion for two-sided procedures that they also provide accurate one-sided statements, at least to the probability matching $O(n^{-1/2})$. While Ghosh argues strongly for the probability matching property, his argument does not seem to take into account the cancellation inherent in (2). We have heard some others argue in favor of such a requirement and some argue against it. We do not wish to take a strong position on this issue now. Perhaps it depends somewhat on the practical context—if in that context the confidence bounds may be interpreted and used in a one-sided fashion as well as the two-sided one, then perhaps probability matching is called for.

I. Ghosh's comments are a reminder that asymptotic theory is useful for this problem, even though exact calculations here are entirely feasible and convenient. But, as Ghosh notes, asymptotic expressions can be startingly accurate for moderate sample sizes. Asymptotics can thus provide valid insights that are not easily drawn from a series of exact calculations. For example, the two-sided intervals also obey an expression analogous to (1),

$$Pr_p(L_n \le p \le U_n) = 1 - \alpha + S_n(p) + Osc_n(p) + O(n^{-3/2}). \tag{3}$$

The term $S_n(p)$ is $O(n^{-1})$ and provides a useful expression for the smooth center of the oscillatory coverage plot. (See Theorem 6 of BCD (2000a) for a precise justification.) The following plot for $n = 50$ compares $S_n(p)$ for five confidence procedures. It shows how the Wilson, Jeffreys and chi-squared likelihood ratio procedures all have coverage that well approximates the nominal value, with Wilson being slightly more conservative than the other two.

As we see it our article articulated three primary goals: to demonstrate unambiguously that the Wald interval performs extremely poorly; to point out that none of the common prescriptions on when the interval is satisfactory are correct and to put forward some recommendations on what is to be used in its place. On the basis of the discussion we feel gratified that we have satisfactorily met the first two of these goals. As Professor Casella notes, the debate about alternatives in this timeless problem will linger on, as it should. We thank the discussants again for a lucid and engaging discussion of a number of relevant issues. We are grateful for the opportunity to have learned so much from these distinguished colleagues.

ADDITIONAL REFERENCES

AGRESTI, A. and CAFFO, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *Amer. Statist.* **54**. To appear.

AITKIN, M., ANDERSON, D., FRANCIS, B. and HINDE, J. (1989). *Statistical Modelling in GLIM*. Oxford Univ. Press.

BOOS, D. D. and HUGHES-OLIVER, J. M. (2000). How large does n have to be for Z and t intervals? *Amer. Statist.* **54** 121–128.

BROWN, L. D., CAI, T. and DASGUPTA, A. (2000a). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* To appear.

BROWN, L. D., CAI, T. and DASGUPTA, A. (2000b). Interval estimation in exponential families. Technical report, Dept. Statistics, Univ. Pennsylvania.

BROWN, L. D. and LI, X. (2001). Confidence intervals for the difference of two binomial proportions. Unpublished manuscript.

CAI, T. (2001). One-sided confidence intervals and hypothesis testing in discrete distributions. Preprint.

COE, P. R. and TAMHANE, A. C. (1993). Exact repeated confidence intervals for Bernoulli parameters in a group sequential clinical trial. *Controlled Clinical Trials* **14** 19–29.

COX, D. R. and REID, N. (1987). Orthogonal parameters and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49** 113–147.

DASGUPTA, A. (2001). Some further results in the binomial interval estimation problem. Preprint.

DATTA, G. S. and GHOSH, M. (1996). On the invariance of noninformative priors. *Ann. Statist.* **24** 141–159.

DUFFY, D. and SANTNER, T. J. (1987). Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* **43** 81–94.

FISHER, R. A. (1956). *Statistical Methods for Scientific Inference*. Oliver and Boyd, Edinburgh.

GART, J. J. (1966). Alternative analyses of contingency tables. *J. Roy. Statist. Soc. Ser. B* **28** 164–179.

GARVAN, C. W. and GHOSH, M. (1997). Noninformative priors for dispersion models. *Biometrika* **84** 976–982.

GHOSH, J. K. (1994). *Higher Order Asymptotics*. IMS, Hayward, CA.

HALL, P. (1983). Chi-squared approximations to the distribution of a sum of independent random variables. *Ann. Statist.* **11** 1028–1036.

JENNISON, C. and TURNBULL, B. W. (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics*, **25** 49–58.

JORGENSEN, B. (1997). *The Theory of Dispersion Models*. CRC Chapman and Hall, London.

LAPLACE, P. S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.

LARSON, H. J. (1974). *Introduction to Probability Theory and Statistical Inference*, 2nd ed. Wiley, New York.

PRATT, J. W. (1961). Length of confidence intervals. *J. Amer. Statist. Assoc.* **56** 549–567.

RINDSKOPF, D. (2000). Letter to the editor. *Amer. Statist.* **54** 88.

RUBIN, D. B. and SCHENKER, N. (1987). Logit-based interval estimation for binomial data using the Jeffreys prior. *Sociological Methodology* **17** 131–144.

STERNE, T. E. (1954). Some remarks on confidence or fiducial limits. *Biometrika* **41** 275–278.

TIBSHIRANI, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76** 604–608.

WELCH, B. L. and PEERS, H. W. (1963). On formula for confidence points based on intergrals of weighted likelihoods. *J. Roy. Statist. Ser. B* **25** 318–329.

YAMAGAMI, S. and SANTNER, T. J. (1993). Invariant small sample confidence intervals for the difference of two success probabilities. *Comm. Statist. Simul. Comput.* **22** 33–59.