10. I regret that I have not had time to do the mathematical work that would, I believe, support some of the above statements.
11. It should also be remembered that the literature on the principle of conditionality is extensive.
12. A general principle, like a mathematical assertion beginning with a universal quantifier, can be refuted by a single counterexample but cannot be validated or proved by any number of special examples.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

REJOINDER[1]

LAWRENCE D. BROWN

Explanations, etc. Several discussants have offered supplementary explanations for the inadmissibility result of Section 3.3 (*Casella*, *Copas*, *Efron*, *Gleser*, *Morris*). Each of the explanations is somewhat different and each adds further understanding.

Gleser goes further and gives a useful extension of my results in the style of my Lemmas 3.3.1 and 3.3.3. Consider the situation discussed in my Section 4.2 where it is desired to estimate the linear function $\kappa = a\alpha + b'\beta$ in the regression setting. Then, if $r \geq 3$, *Gleser*'s Theorem 1 can be applied via his formula (5) to yield a specific, useful estimator dominating $\delta_0 = a\hat{\alpha} + b'\hat{\beta}$. The existence of a dominating estimator was already established in my Theorem 4.2, even for $r = 2$, but no usable formula was given.

Lu demonstrates that the general nature of the inadmissibility phenomena here is not significantly dependent on the form of the loss function. Insofar as his results for L_c are not directly constructive (analogous to my Theorems 2.2.1 and 3.2.2 for squared error) they point to the important question of constructing estimators in the regression setting which usefully dominate δ_0 under L_c.

Limited translation estimators. *Morris* (explicitly) and *Efron* (implicitly) each raise the issue of modifying the proposed estimators to limit maximum coordinatewise risk. (This appears to be the joint occurrence of conditionally independent but marginally highly correlated events!) *Berger* also makes this suggestion. This seems reasonable, particularly in view of the numerical results *Berger* mentions. However it is important to understand the justification for this suggestion before putting it into practice.

To do so consider the usual multiple normal means estimation problem and the positive-part James–Stein estimator, which is given by d^*_+ of (2.1.7) for $\Sigma = \Omega = I$ and $\rho = p - 2$. For moderate $p \geq 3$ this is known to approximate a

[1]Research supported in part by NSF DMS 88-09016.

generalized Bayes estimator for a prior which has—as *Berger* writes—"flat... but not too flat" tails. Let us assume that this is also the case for large p. [Berger (1985, page 238) informally claims this is so, but I am still skeptical in the current case. If it is not so, then an appropriate generalized Bayes estimator should replace it in the following discussion.] In this case the limited translation estimator will not dominate d_+^*. But there are two kinds of robustness arguments which nevertheless can justify use of the limited translation estimator. One involves robustness in the loss function. If there is a possibility that some coordinate(s) of the loss are more important than the others, then it becomes desirable to limit the maximum coordinatewise risk. [See Brown (1975) for a precise formulation, but the idea is already implicit in Efron and Morris (1972).] Alternatively, if there is some doubt as to the suitability of the generalized prior, then certain types of robust Bayesian considerations suggest using something like the limited translation estimators in place of d_+^* [See, e.g. Berger (1985), pages 243–244, Dey and Berger (1983) and Berger and Dey (1985)].

Now consider the current problem of estimating α when V is ancillary. Here the first type of robustness motivation cannot exist since the loss is one-dimensional. The limited translation alternatives must be appealing—if at all—because they are more appropriate for a realistic, robust prior assessment than are those in Section 3.2. It is thus not the large value of the conditional risk, per se, that justifies the limited translation estimator; Rather it is that this large value is a warning that the prior may not be realistic or, in other words, that what Berger (1984) has called "posterior robustness" does not hold. Perhaps it is also of interest to note that essentially this robust Bayesian idea is already apparent in Efron and Morris (1971), which proposed limited translation estimators for the standard one-dimensional normal problem.

An additional comment seems relevant. *Gleser* has pointed out how it is possible in the regression-with-ancillary setting to construct an estimator which simultaneously improves on all coordinates of the estimated (α, β') vector. It is plausible that, even here, it would be desirable to construct a limited translation version if a suitable version could be constructed and shown to have some meaningful, desirable robustness properties.

Conditional criteria versus marginal criteria, I. *Berger* contains an incisive analysis of the conditionality issues raised in the current situation. I agree with almost all he has to say. His "alternative ancillarity paradox" is particulary pertinent. He may also be right when he later continues that "it is rare for estimators developed *solely* [my italics] as unconditional frequentist dominating estimators to be of much use in practice." In any case, I would not recommend proceeding in such an exclusive fashion. Of course, "developing the procedure conditionally to assure conditional soundness and (only then) checking its unconditional behavior," as he suggests, can also mislead. The ancillarity paradox in my paper is meant to be an example. Pandora or no (I am not sure which) I agree with him that "a mix of good conditional and good unconditional frequentist behavior is needed."

Conditional versus marginal criteria, II. When *Berger* suggests first analyzing problems conditionally, he is careful to make clear that this conditional analysis should be Bayesian in some sense. Furthermore, as we both agree, the prior (or robust family of priors) must be chosen without regard to the observed value of the ancillary. To proceed conditionally in some other fashion is much more dangerous. *Fraser and Reid* seem to recognize this in Section 5 of their discussion. There they present an example in which use of conditionally minimax estimators leads to a marginally unacceptable, nonminimax procedure. A systematic treatment of a much more important general instance of this phenomenon can be found in He (1989). [*Fraser and Reid*'s example is perhaps not so "simple" as they thought. They have overlooked the fact that the conditional problem inherits the parameter space of the marginal one. Thus, given $x_{12} + x_{22} = 1$, the variable x_{22} is Bernoulli with $\frac{1}{2} \le p = (1 + \theta)/2 \le 1$ since $0 \le \theta \le 1$. The admissible minimax procedure for this conditional problem is $(2 - \sqrt{2})x_{22} \approx 0.5858x_{22}$. This has maximum risk (over $0 \le \theta \le 1$) of 0.1716 attained at $\theta = 0$ and 1. Rao–Blackwellizing this conditional estimator together with that for $x_{11} + x_{21} = 1$ does indeed improve over the simple conditional prescription. However, it does not improve the maximum risk at $\theta = 1$ since $P_{\theta=1}(x_{12} = x_{21} = 0) = 1$. Thus both the fully conditional original and the Rao–Blackwellized estimators have maximum risk $\frac{1}{3} \times \frac{1}{16} + \frac{2}{3} \times 0.1716 = 0.1352$. The true minimax procedure for the marginal problem is $\tilde{t}_a^* = x_{11} + (2 - \sqrt{2})x_{22}$, which is Bayes for a suitable prior supported on $\theta = 0, 1$. This procedure has maximum risk of only 0.1144, attained at $\theta = 0, 1$.] Given their discussion, and particularly this example, I do not understand their conclusion that "the issue of marginal optimality is not of interest. . . ." Perhaps they are thinking only of situations similar to Welch's confidence interval example, which I will now discuss.

Welch's and Cox's examples. The discussants are very divided here. *Fraser and Reid* strongly support the standard (conditional) analysis in the Welch example, and presumably also in the Cox example. *Efron* supports the standard analysis in Cox's example, but for what seems to me an unusual reason. *Stein* [in his (7) and (8)] also supports the conditional analysis in the Welch example but favors the marginal test. *Berger* favors the marginal test in the Cox example. [It seems like a contradiction for the avowed frequentist (*Stein*) to prefer the conditional test and the avowed Bayesian (*Berger*) to prefer the marginal one! But read them closely.] *Gleser* is the only discussant who seems to have understood what I was hinting in my Section 5 so I will be explicit here—but also brief.

What I propose as a level $\alpha = 0.05$ procedure in Cox's example is this: If the smaller sample size n_1 obtains, then use the formally level 0.1 conditional test *and state that the test used has (conditional) level* 0.1. Otherwise (n_2) use the appropriate level ε (ε near 0) conditional test *and state that the test used has conditional level* ε.

My suggestion is thus that the standard test statement consisting of α plus either "accept" or "reject" be augmented with a statement of the conditional level. A formal framework for this sort of thing can be found in Brown (1978); see also Kiefer (1976, 1977). Within this framework the above proposal dominates the ordinary conditionally level 0.05 test and I suspect it is admissible. Note also that this proposal follows—and in a way justifies—a customary practice of using smaller α levels for higher quality data. My suggestion for Welch's example is similar.

In accordance with the above I think it would be very useful to have available—as *Gleser* suggests—supplementary estimates of the conditional risk of the procedures in Section 3.2. Perhaps this would also help settle the valid confusion expressed in *Morris*'s Section 1. *Johnstone* makes a useful start in this direction.

Frequentist measures of performance. I think *Gleser*'s description is overly restrictive. He writes that "frequentist measures... implicitly assume a 'stream' of *similar* [my italics] experiments... ." (Some other writers even require identical experiments. *Hill* comes close to this in his third-from-the-end paragraph.) In fact the experiments need not be similar at all. For my view on this, see my discussion of Berger (1984, pages 126–127), only part of which is cited by *Hill*.

The real world is finite. *Hill* argues that the conflict between frequentist admissibility (FA) and the ancillarity principle (AP) "cannot occur in the case of parameters and data for which there are only a finite number of possibilities." Make the proviso that this finite list of possibilities be known to the statistician. Then *Hill* is correct in the strictly formal sense that every admissible procedure is (stepwise) Bayes both marginally and conditionally. [See Hsuan (1979) for the definition of stepwise Bayes.] As *Berger* has noted in his discussion, the only conflict between FA and AP that could then occur would be if different (stepwise) priors were to be used in different conditional problems. Nevertheless, I disagree with *Hill*.

There are two related practical faults in this reasoning. One is that the preceding proviso is often not present in any realistic fashion. The second is that there is a hidden presumption that the unique, suitable prior can be determined. *Hill* seems to argue that when this prior cannot otherwise be determined, then it suffices to simply impose a uniform prior on the finite parameter space.

To illustrate these faults I will consider below a slightly whimsical example. Also, note that because of the mathematical connections it suffices here to look at the ordinary simultaneous-estimation-of-normal-means problem. (See especially the explanation in my Lemma 3.3.1 and in *Efron*'s comment.)

So, let us suppose that for four different, previously unnoticed species of bees, living on four different continents, we wish to estimate the average species-wide log-mass per beehive. Let us further assume that previous experi-

ence convinces us that a suitable approximation is to assume, that the observed log-mass x_i is a $N(\mu_i, \sigma_i^2)$ variable. For simplicity in the current discussion let us also assume that σ_i^2 is known and that $\sigma_i^2 \equiv 1$. (Neither of these assumptions is needed for the general assertions that follow if several beehives for each species can be measured.)

Hill would begin by noting that both X and μ are really discrete; let us say in multiples of $1/1000$. (This assumption, even if questionable, does not materially affect the argument to follow.) Accordingly, the sampling distribution now becomes a discretized normal.

More importantly, he would argue (correctly) that both the μ_i and X_i are bounded; let us say $b \leq \mu_i, X_i \leq B$. The dilemma lies in the values of b, B. I think we can agree that it suffices (i.e., is not too small) to choose B equal to the log mass of a herd of elephants. An analogously small value could be agreed upon for b. We have now arrived at a strictly finite problem in which the Bayes procedures are a complete class. But that does not help settle the problem of what estimator to use.

Hill would now, it seems, have us use the Bayes procedure for the uniform prior on the possible values of $\mu = (\mu_1, \ldots, \mu_4)$ with each μ_i between b and B. This Bayes procedure will be exactly X so long as $b \ll X_i \ll B$. [Elsewhere in his discussion he seems to argue for use of the least squares estimator, by which I presume he means either X itself or the truncated version with coordinates $\min(\max(X_i, b), B)$. In either case the least squares estimator is inadmissible for the truncated problem, and so does not seem to meet the test of his Theorem 1.]

The discretized positive part James–Stein estimator centered on a prior guess for μ would be preferable. An alternative if this prior guess for μ were vague would be to use the corresponding Lindley–Smith James–Stein positive part estimator shrinking toward $\overline{X}(1, \ldots, 1)$. Both of these estimates would be only approximately admissible. If an exactly admissible estimator were desired, the proper approach would be to use the Bayes procedure for a truncated and discretized version of a suitable prior, like those suggested in Strawderman (1971) or Berger (1976, 1984, 1985). The result should be similar to the positive part James–Stein estimator suggested above.

These James–Stein style estimators will dominate the uniform prior Bayes estimator in risk except when some μ_i is near b or B. But values of μ_i near either extreme are unrealistic. No one a priori expects the beehive to weigh as much as an elephant, and if it did, then we might seriously question our assumption that $X_i \sim N(\mu_i, \sigma_i^2)$, as well as our sanity and safety!

In summary, the James–Stein estimator (or something much like it) often remains a preferable alternative to the uniform prior Bayes estimator even in a finite world, since the bounds on this world are frequently more extreme than reasonable prior opinion. Berger and Wolpert (1988, page 189) make a similar point. Because of this, the apparent conflict persists in practice between what *Hill* terms FA and AP. A final quotation from *Hill*, with which I do agree, is appropriate. Elsewhere [Hill (1981)] he wrote: "Even if we accept that real problems are finite [infinite and] continuous idealizations are com-

monly made in statistics for practical approximations, so the question would remain as to when such idealizations are dangerous."

To conclude. I would like to thank the discussants for their stimulating and penetrating discussions. I am sorry that it has been impossible here to respond in some way to all of the interesting issues they have raised. I also wish to thank the editors of the *Annals* for encouraging this discussion. In spite of the risk of being repetitive, it nevertheless seems appropriate to close with an evaluation taken from Efron (1982). Efron was referring to the usual James–Stein situation, but his advice is equally relevant in the ancillary situation presented here: "A successful answer is likely to be at least partly Bayesian while still enjoying good frequentist properties."

REFERENCES

BERGER, J. O. (1984). The robust Bayesian viewpoint (with discussion), In *Robustness of Bayesian Analysis* (J. Kadane, ed.) 321–372. North-Holland, Amsterdam.

BERGER, J. O. and WOLPERT, R. L. (1988). *The Likelihood Principle*, 2nd ed. IMS, Hayward, Calif.

BROWN, L. D. (1975). Estimation with incompletely specified loss functions. *J. Amer. Statist. Assoc.* **70** 417–427.

HE, KUN (1989). A paradox in the estimation of multinomial probabilities. *J. Amer. Statist. Assoc.* To appear.

HILL, B. M. (1980). On some statistical paradoxes and nonconglomerability (with discussion). In *Bayesian Statistics* (J. M. Bernardo, et. al., eds.) 39–66. Univ. Press, Valencia, Spain.

HSUAN, F. C. Y. (1979). A stepwise Bayesian procedure. *Ann. Statist.* **7** 860–868.

EFRON, B. (1982). Maximum likelihood and decision theory. *Ann. Statist.* **10** 340–356.

EFRON, B. and MORRIS, C. (1971). Limiting the risk of Bayes and empirical Bayes estimators, I: The Bayes case. *J. Amer. Statist. Assoc.* **66** 801–815.

EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators, II: The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139.