Comment

Robert E. Bechhofer and Ajit C. Tamhane

We had little imagined that BTIB designs for comparing test treatments with a control treatment would generate such a wide research interest among the design theorist community when we first proposed this new class of designs in Bechhofer and Tamhane (1981). Naturally we are very pleased and gratified to note the tremendous progress that has been made in the last seven years in the study of these designs and their extensions, with particular emphasis on the optimality question. Hedayat, Jacroux and Majumdar, who have been at the forefront of this research, are to be congratulated for providing a fine survey of the results. We thank the Editor for giving us an opportunity to discuss this survey.

Hedayat, Jacroux and Majumdar focus their attention on A- and MV-optimal designs. Both of these optimality criteria refer to minimizing suitable functions of the variances of the $\hat{t}_i - \hat{t}_0$, but do not take their *correlations* into account. (We follow the same notation as in their article.) Thus the optimal designs derived would seem to be appropriate when the results of the experiment are to be reported in terms of the above *point* estimates accompanied by their estimated standard errors or in terms of *separate* confidence interval estimates of the $t_i - t_0$, $i = 1, \cdots, v$. However, in many applications a *simultaneous* confidence region (or a set of simultaneous confidence intervals) is more appropriate than separate confidence intervals. The following is an example of such an application. It is frequently desired to select one or more of the test treatments for eventual use. The primary selection criterion is the parameter $t_i - t_0$ (test treatments with large values being preferred, say) but there also are secondary criteria such as costs. The precise rules for selection of the test treatments cannot be stated in advance. For example, depending on the experimental results and other side considerations, the two apparently "best" test treatments (in terms of the $\hat{t}_i - \hat{t}_0$ values) may be selected or even the third apparently "best" test treatment may be selected. A set of *simultaneous* confidence intervals guarantees a specified confidence coefficient regardless of which test treatments are selected and for which the corresponding confidence interval estimates are reported. For other examples of applications where simultaneous inference (multiple comparisons) procedures are called for, see Hochberg and Tamhane (1987, Chapter 1).

Robert E. Bechhofer is Professor, School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York 14853. Ajit C. Tamhane is Professor, Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, Illinois 60201 and is also a faculty member in the Department of Industrial Engineering and Management Sciences, Technological Institute, Northwestern University.

Under normal theory, operating characteristics of simultaneous inference procedures are generally functions of not only the variances of the $\hat{t}_i - \hat{t}_0$ but also their correlations. It is true that, for example, A-optimality is equivalent to minimizing the sum of the lengths of the axes of the *simultaneous* confidence ellipsoid (assuming the usual normal theory model) for the given contrasts of interest. But curiously, D-optimality, which corresponds to minimizing the volume of the confidence ellipsoid, and which does take into account the variances as well as the correlations, is not a useful optimality criterion for the present problem, as the authors rightly point out.

We believe that the use of these standard optimality criteria due to Kiefer (1958) is questionable in experiments involving multiple comparisons of test treatments with a control because they do not address the typical inferential goals in such experiments. The authors state that "To begin with we need to postulate a model ⋯" In the same vein, it is also true that, to decide on a design (optimal or efficient), we need to postulate the types of *inferences* that will be made based on the data collected from the experiment. The authors make a brief reference to this point when they state that "⋯ our primary goal is to determine which among the test treatments might be better than the control ⋯" However, we do not think that this goal necessarily translates into "⋯ to estimate the magnitude of each $t_i - t_0$ with as much precision as possible" without reference to how the resulting estimates will be used to determine the apparently better test treatments. In fact, as we explain below, two types of inferential goals are appropriate in these experiments, and both involve taking into account the variances of the $\hat{t}_i - \hat{t}_0$ as well as their correlations.

Often, in exploratory stages of an investigation there are a large number of new candidate test treatments, and the goal is to screen out those that are inferior to the control treatment. For this goal the subset selection formulation of Gupta and Sobel (1958) would appear to be suitable. The test treatments that are selected in this initial experiment can then be studied more intensively in the confirmatory stage. Moreover, it usually is required (e.g., by a regulatory agency such as the Food and Drug Administration) that the control also be included in this stage. The goal of this second experiment is to obtain precise confidence interval estimates of the control versus test treatment contrasts (and also possibly pairwise treatment contrasts). For this goal the simultaneous confidence estimation approach proposed by Dunnett (1955) would appear to be suitable.

It is important to note that the simultaneous confidence region provided by Dunnett's procedure is "*rectangular*" in shape, not ellipsoidal. The rectangular confidence region is more useful and relevant in practice for the present problem because (a) it is easier to interpret, and (b) the ellipsoidal confidence region gives much longer confidence intervals when it is projected onto the $t_i - t_0$ axes. This is so because an ellipsoidal confidence region is optimal when *all* linear functions of the $t_i - t_0$ for $i = 1, \cdots, v$ are *a priori* of interest; thus, it performs conservatively for the specific simple functions, namely, the $t_i - t_0$, which are the only ones of interest in the present problem. The second point to note is that the operating characteristics of the Gupta-Sobel subset selection procedure (e.g., its probability of correct selection) and the Dunnett simultaneous confidence estimation procedure (e.g., its joint coverage probability) depend on the variance of the $\hat{t}_i - \hat{t}_0$ as well as on their correlations. In fact, the Gupta-Sobel subset selection procedure and the Dunnett procedure for one-sided simultaneous confidence intervals are very closely related, both being based on the same one-sided multivariate Student t percentage point; the two-sided Dunnett procedure is based on the corresponding two-sided percentage point.

It was with the above background that one of us (Bechhofer) was motivated to study the problem of optimal allocation of observations for zero-way elimination of heterogeneity designs (completely randomized designs). Dunnett (1955) had shown numerically for his rectangular simultaneous confidence region that the $\sqrt{v}$ allocation rule, i.e., $r_{d0} = \sqrt{v}\, r_{d1}$ and $r_{d1} = \cdots = r_{dv}$, is approximately optimal (in the sense of maximizing the joint confidence coefficient for a fixed total sample size n) for large values of n, i.e., for large values of the joint confidence coefficient, $1 - \alpha$. Bechhofer (1969) also used the criterion of maximizing the joint confidence coefficient for given n, and derived the optimal allocation (using a continuous approximation to the sample sizes) for one-sided simultaneous confidence intervals of the form $\{t_i - t_0 \leqq \hat{t}_i - \hat{t}_0 + a (1 \leqq i \leqq v)\}$ for specified "allowance" a and for any value of $1 - \alpha$. He also showed that asymptotically (as $n \to \infty$) the $\sqrt{v}$ allocation rule is optimal. These results were extended to two-sided simultaneous confidence intervals of the form $\{t_i - t_0 \varepsilon [\hat{t}_i - \hat{t}_0 + a] (1 \leqq i \leqq v)\}$ by Bechhofer and Nocturne (1972).

Now the asymptotically optimal $\sqrt{v}$ allocation rule corresponds to the A- and MV-optimality criteria (if the integer restrictions on the sample sizes are ignored). Therefore these criteria would seem to apply to the simultaneous confidence estimation problem *only for large sample sizes* (although it is true that the approach to the asymptotically optimal allocation is fairly rapid as can be seen in the tables given in Bechhofer and Tamhane (1983a)). However, most of the work in optimal designs is concerned with small sample sizes. This is particularly true for the elimination of heterogeneity designs (e.g., block designs) with which the Hedayat, Jacroux and Majumdar article is principally concerned. We now turn our discussion to these designs.

In our studies we focused on one-way elimination of heterogeneity where the block size k is less than the total number of treatments, $v + 1$. Based on symmetry considerations, we proposed a new class of so-called BTIB designs, which have the following statistical balance property:

$$\mathrm{var}(\hat{t}_i - \hat{t}_0) = \tau^2\sigma^2, \quad 1 \leqq i \leqq v,$$

and

$$\mathrm{corr}(\hat{t}_i - \hat{t}_0, \hat{t}_{i'}, - \hat{t}_0) = \rho,$$

$$i \neq i', 1 \leqq i, i' \leqq v;$$

here the parameters τ^2 and ρ depend on the design and σ^2 denotes the common error variance. This statistical balance property is equivalent to the combinatorial balance property given in Definition 2.2 of the Hedayat, Jacroux and Majumdar article (see Theorem 3.1 in Bechhofer and Tamhane, 1981). We called these designs BTIB because they are balanced with respect to the test treatments. We next addressed the problem of finding an optimal BTIB design, which for given v and k, and for specified standardized "allowance" a/σ and joint (one-sided or two-sided) confidence coefficient $1 - \alpha$, requires the smallest possible number of blocks, b. In the search for an optimal design we could eliminate any *inadmissible* design, which gives a lower joint confidence coefficient for *all* values of a/σ than another design with no larger b. A design that is not inadmissible is called *admissible*. We characterized inadmissible designs by the following simple rule (see Theorem 5.1 in Bechhofer and Tamhane (1981)): For given k and v, a BTIB design d' with parameters b', τ'^2 and ρ' is inadmissible with respect to another BTIB design d with parameters b, τ^2 and ρ if $b \leqq b'$, $\tau^2 \leqq \tau'^2$ and $\rho \geqq \rho'$ with at least one strict inequality. This rule is based on the fact that, under normality, the joint confidence coefficient (one-sided or two-sided) is a decreasing function of τ^2 and an increasing function of ρ.

Examples of A-optimal designs that are not optimal for our simultaneous confidence interval estimation criterion are easy to find. For example, the design consisting of 6 copies of

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 2 \\ 2 & 3 & 3 \end{bmatrix}$$

is given in Table 1 of Hedayat and Majumdar (1984) as A-optimal among *all* designs for $k = 3$, $v = 3$ and $b = 18$. From Table OPT1.3.3 of Bechhofer and Tamhane (1985) it is seen that this design is *not* optimal even in the restricted class of BTIB designs for $1 - \alpha \leqq 0.7653$ (but is optimal for larger values of $1 - \alpha$). Many more such examples can be found. This is not very surprising, of course, because the two criteria are different. We recognize that different criteria can lead to different optimal designs, and thus it is unfair to judge optimal designs based on one criterion relative to the other. Furthermore, admittedly our criterion is based on the normality assumption, whereas the authors' criteria are not based on any particular distribution.

However, our admissibility criterion, although also derived from the joint coverage probability calculation under the normality assumption, is much weaker. In other words, if a BTIB design d requires no more blocks b, and yet yields no larger variance $\tau^2\sigma^2$ and no smaller correlation coefficient ρ than another BTIB design d' then, in general, the latter should not be used. We were surprised to find that several of the A-optimal BTIB designs given in Hedayat and Majumdar (1984) are inadmissible. In Table 3 of their paper we point out three examples of A-optimal designs that are inadmissible: For $k = 2$ and $v = 3$ let

$$d_0 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 2 & 3 \end{bmatrix} \quad \text{and} \quad d_1 = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 3 & 3 \end{bmatrix}.$$

Then for $b = 6$ the design $2d_0$ with $\tau^2 = 1$ and $\rho = 0$ is inadmissible with respect to the design $d_0 \cup d_1$ (i.e., d_0 union with d_1) which has $\tau^2 \doteq 1$ and $\rho = 0.5$. Similarly for $b = 18$ the design $5d_0 \cup d_1$ which has $\tau^2 = 0.3$ and $\rho = 1/6$ is inadmissible with respect to the design $4d_0 \cup 2d_1$ which has $\tau^2 = 0.3$ and $\rho = 1/3$. One might say that in each of these two examples both of the competing designs are A-optimal, and Hedayat and Majumdar's algorithm identified the one that unfortunately had the smaller ρ-value. However, we next give an example where this is not the case. The BTIB design consisting of five copies of

$$d_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{bmatrix}$$

with $\tau^2 = 0.4$ and $\rho = 0$ is given as A-optimal among all BTIB designs for $k = 2$, $v = 6$ and $b = 30$. However, the design $2d_0 \cup d_1$ with $b = 27$, $\tau^2 = 0.3750$ and $\rho = 1/3$ is superior on *all three counts*, and hence $5d_0$ is inadmissible; here

$$d_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 5 \\ 2 & 3 & 4 & 5 & 6 & 3 & 4 & 5 & 6 & 4 & 5 & 6 & 5 & 6 & 6 \end{bmatrix}.$$

The design $2d_0 \cup d_1$ is given as A-optimal for $b = 27$ in Hedayat and Majumdar's (1984) Table 3. But because their algorithm did not compare designs for different b's, it failed to note that this design is actually superior (even in terms of the A-optimality criterion) to the A-optimal design for the next higher b, namely $b = 30$.

Our admissibility criterion has certain other desirable features, which permit one to restrict consideration for given k and v to the so-called *minimal complete set of generator designs* for constructing any BTIB design for that (k, v). The members of the minimal complete set serve as building blocks for larger BTIB designs. Such minimal complete sets were constructed for selected (k, v)-combinations by Notz and Tamhane (1983) and Türe (1982). For $k = 2$, $v \geqq 2$ and for $k = 3$, $v = 3$ it is easy to see that the minimal complete set consists of just two generator designs. In this case the analysis required to determine the admissible and optimal designs is considerably less difficult and is given in Bechhofer and Tamhane (1983b). It may be of interest to note that the result given in equation (3.11) of that article for characterizing admissible BTIB designs for $k = 2$ is the same as that given in Theorem 3.1 of Hedayat and Majumdar (1984) for characterizing A-optimal BTIB designs.

We conclude our discussion by noting some of the problem areas that need further work. The first on our list is the designs for two-way elimination of heterogeneity. Much remains to be done in this area, particularly on the problem of constructing "balanced with respect to test treatments row-column designs" (analogous to BTIB designs for one-way elimination of heterogeneity). The problem of determining the minimal complete set of generator designs for this case is an important one, but quite likely a difficult one. Presumably these designs could be derived from Latin squares, Youden squares and perhaps lattice designs (for large v). Some ad hoc construction methods have been given in a Ph.D. dissertation at Virginia Polytechnic Institute by Rashed (1984).

The next problem on our list is that of finding good designs for comparing test treatments with several controls. As noted by the authors, a beginning has been made in this area of research. In future work it would be desirable to keep some important practical features of the problem in mind. One such feature is that the comparisons with different controls may not be required to be of equal precision. For example, in a clinical trial for a new drug it is not uncommon to include two controls, a placebo and an existing active drug. For regulatory purposes, it often is necessary to demonstrate the magnitude of the activity of the new drug, and therefore the comparison with the placebo is the more important. It is not always necessary to demonstrate to the regulatory agency that the new drug is more effective than the existing drug. But for the purposes of the pharmaceutical company's marketing efforts, in fact, the second comparison is likely to be the more important. This latter comparison would generally be two-sided. Such considerations should be taken into account before determining how to optimally allocate the available experimental resources to different competing test treatments and the controls.

A final brief note concerning nomenclature. We suggest that the word "control" should be used rather than "standard" because the latter sometimes refers to a *known* benchmark value; this is the case, e.g., in the physical sciences (although, not always in the biological sciences). Clearly, if the comparisons are made with a known benchmark then the device of blocking cannot be used.

We again express our gratitude to the authors for this state-of-the-art survey and to the editor for giving us an opportunity to comment on it.

ACKNOWLEDGMENTS

This research was partially supported by the United States Army Research Office through the Mathematical Sciences Institute of Cornell University. We are indebted to Professor Charles Dunnett for his insightful comments and suggestions on an earlier draft of this article.

ADDITIONAL REFERENCES

BECHHOFER, R. E. (1969). Optimal allocation of observations when comparing several treatments with a control. In *Multivariate Analysis II* (P. R. Krishnaiah, ed.) 463–473. Academic, New York.

BECHHOFER, R. E. and NOCTURNE, D. J. M. (1972). Optimal allocation of observations when comparing several treatments with a control. II: 2-sided comparisons. *Technometrics* **14** 423–436.

BECHHOFER, R. E. and TAMHANE, A. C. (1983a). Design of experiments for comparing treatments with a control: Tables of optimal allocations of observations. *Technometrics* **25** 87–95.

BECHHOFER, R. E. and TAMHANE, A. C. (1983b). Incomplete block designs for comparing treatments with a control (II): Optimal designs for one-sided comparisons when $p = 2(1)6$, $k = 2$ and $p = 3$, $k = 3$. *Sankhyā Ser. B* **45** 193–224.

BECHHOFER, R. E. and TAMHANE, A. C. (1985). Tables of admissible and optimal balanced treatment incomplete block (BTIB) designs for comparing treatments with a control. *Selected Tables Math. Statist.* **8** 41–139. Amer. Math. Soc., Providence, R. I.

DUNNETT, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Statist. Assoc.* **50** 1096–1121.

GUPTA, S. S. and SOBEL, M. (1958). On selecting a subset which contains all populations better than a standard. *Ann. Math. Statist.* **29** 235–244.

HOCHBERG, Y. and TAMHANE, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

RASHED, D. H. (1984). Designs for multiple comparisons of control versus test treatments. Ph.D. dissertation, Dept. Statistics, Virginia Polytechnic Institute and State Univ.

Comment

William I. Notz

Sam Hedayat, Mike Jacroux, and Dibyen Majumdar are to be congratulated on this very thorough survey of optimal designs for comparing test treatments with a control. This paper is an excellent starting point for anyone wishing to do research in this area and it is a nice reference for those of us actively engaged in such research. Unfortunately, any such survey begins to go out of date the moment it is completed as research goes ever forward. The authors can do nothing about that, however.

William I. Notz is Associate Professor, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210.

Let me begin my comments by describing the history of my own involvement in this area of research. If nothing else, this will at least add a little historical color.

I first became acquainted with this area of research as a relatively new assistant professor at Purdue. In the Autumn of 1980, Bob Bechhofer came to Purdue as a colloquium speaker. He spoke about results he and Ajit Tamhane had obtained on incomplete block designs for comparing test treatments with a control and which were soon to appear in Bechhofer and Tamhane (1981). One unsolved aspect of the research, which Bob invited those of us in the audience to try and solve, involved constructing finite sets of designs (so-called minimal complete sets of generator designs)