Recent Developments in Multiple Comparison Procedures
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 47 (2004) 1–23

Implementing Shaffer's multiple comparison procedure for a large number of groups

John R. Donoghue[1]

Educational Testing Service

Abstract: Shaffer (1986) presented two more powerful modifications (commonly referred to as S1 and S2) of Holm's (1979) sequentially rejective Bonferroni procedure. Unfortunately, use of the more-powerful S2 procedure has been severely limited by the complexity of its implementation. This paper presents a method of allowing Shaffer's S2 procedure to be used in a much larger class of problems.

Theoretical results concerning an aspect of Shaffer's S2 procedure are derived. Use of these results in implementing the procedure is described. Next, two heuristics are described which greatly enhance the efficiency of the method. Finally, an efficient algorithmic method of implementing Shaffer's procedure is outlined.

The present work allows Shaffer's procedure for multiple comparisons to be applied to a large number of groups. Unlike many other methods, no assumptions about the joint distributions of the test statistics need be made. Shaffer's S2 method and other methods are illustrated by application to two data sets; comparisons of the order of 11 clustering methods, and comparison of the means of 44 states from the 1994 National Assessment of Educational Progress Trial State Assessment of Reading at Grade 4.

1. Introduction

In performing multiple statistical tests, it may be important to control the probability of Type I error over the set (family) of related comparisons. The probability of incorrectly rejecting one or more null hypothesis is termed the familywise error rate (FWER). If each comparison is performed at the nominal significance level α, the FWER can greatly exceed α. Controlling the overall Type I error rate is a standard goal of multiple comparison procedures.

This paper examines the special case of examining all pairwise comparisons of a number of groups, while maintaining the FWER at the nominal level α. Specifically, a method is developed to implement Shaffer's (1986) modification to the Holm (1979) procedure.[2] Although Shaffer's S2 procedure has achieved some use in the past two decades, its use has been limited by the complexity in implementing the procedure. This paper presents a method of managing that complexity, allowing Shaffer's procedure to be used in a much larger class of problems.

[1]Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA. e-mail: jdonoghue@ets.org

Keywords and phrases: Shaffer's procedure, multiple comparisons, familywise error rate, pairwise comparisons, National Assessment of Educational Progress, NAEP.

AMS 2000 subject classifications: primary 62J15.

[2]Shaffer's (1986) paper actually presentes two methods, commonly referred to as S1 and S2 (e.g., Holland & Copenhaver, 1988; Rasmussen, 1993). Unless otherwise indicated, this paper will be concerned with the S2 procedure. See Section 1.2.1 for more details.

1.1. Background

A wealth of multiple comparison procedures is available; no attempt will be made to summarize them here. The interested reader is referred to any of several excellent recent reviews of procedures available (e.g., Hochberg & Tamhane, 1987; Hsu, 1996; Shaffer, 1995; Toothaker, 1991; Westfall, Tobias, Rom, Wolfinger, & Hochberg, 1999).

Two observations help place the present paper within this extensive literature. First, some multiple comparison procedures, such as the early Honestly Significant Difference method (Tukey, 1949) and other early multiple range methods for pairwise comparisons, are based on distribution assumptions such as multivariate normality of observations. Other, more recent methods are based on the Simes inequality, which also requires some assumptions about the joint distribution of the test statistics. As discussed below, Shaffer's procedure does not require such assumptions; only the probabilities of the individual statistical tests, and the structure of the tests, e.g., all pairwise comparisons, are needed.

A fundamentally different approach is control of the False Discovery Rate FDR (Benjamini & Hochberg, 1995). This approach abandons control of the familywise Type I error. Instead, it limits the expected proportion of false rejections among the rejected hypotheses, while Shaffer's procedure is designed to control the more stringent FWER. Clearly, the choice between controlling the FWER and the FDR depends upon the specific testing situation.

1.1.1. Bonferroni inequality

Probably the earliest (and still widely used) multiple comparison procedure is based on the Bonferroni inequality, which states that the overall probability of at least one false rejection of the null hypothesis (i.e., the FWER) is less than or equal to the sum of the probabilities of the individual events. Thus, to control the familywise Type I error at $\alpha$of a set of R comparisons, each test is performed at $\alpha' = \alpha/R$.

The Bonferroni logic has the advantages that it is simple to apply and is applicable across a wide variety of situations. However, for even a moderate number of comparisons R, the Bonferroni method becomes very conservative. There are modifications to the Bonferroni procedure that enhance the power of the procedure while still maintaining FWER control. The modifications examined in this paper make use of a sequential logic to control the Type I error rate.

1.2. Modifications to the Bonferroni–Sequential procedures

In sequentially rejective testing procedures (e.g., Holm, 1979; Shaffer, 1986; Hommel, 1988; Hochberg, 1988), the p-values associated with the R tests are sorted from smallest to largest ($P_1 \leq P_2 \leq \cdots \leq P_R$), with the associated hypotheses ordered identically $H_1, \ldots, H_R$. The critical probability to be used is a function of the comparison; P_ℓ is tested against α_ℓ. For the unmodified Bonferroni procedure, $\alpha_\ell \equiv \alpha/R$. Sequentially rejective procedures gain power over the unmodified Bonferroni procedure by modifying α_ℓ for each ℓ. FWER is maintained by the closed nature of the procedure (Marcus, Peritz, & Gabriel, 1976; Grechanovsky & Hochberg, 1999). So-called step-down procedures, such as Holm (1979) or Shaffer (1986), maintain FWER by testing step $\ell+1$ only if the hypothesis associated with step$\ell$was rejected (i.e., $P_\ell \leq \alpha_\ell$). Step-up procedures (e.g., Hochberg, 1988; Dunnett & Tamhane, 1992) proceed in the opposite fashion, first testing P_R against α_R and proceeding to $H_{\ell-1}$ only if H_ℓ was not rejected. See Liu (1996), Grechanovsky and

Hochberg (1999), and Finner and Roters (2002), for recent reviews of step-down and step-up procedures.

Holm (1979) proposed a sequentially-rejective Bonferroni procedure, the earliest proposed generally-applicable sequential procedure. Holm's step-down procedure tests P_1 against $\alpha_1 = \alpha/R$. If the associated hypothesis is rejected, test P_2 at $\alpha_2 = \alpha/(R-1)$, and so on, testing P_ℓ at $\alpha_\ell = \alpha/(R-\ell+1)$ until one hypothesis $H_\ell (1 \leq \ell \leq R)$ is not rejected, at which point testing stops. The hypotheses associated with tests $1, 2, \ldots, \ell-1$ are declared significant, while hypotheses H_ℓ through H_R are not rejected (see Westfall & Young, 1993, pp. 72–74 for a theoretical discussion).

1.2.1. Shaffer's modified Bonferroni procedure (step-down)

Shaffer (1986) pointed out that it is often possible to make use of the logical interrelationships among the hypotheses to reduce the family size in successive steps beyond that in the Holm procedure. Consider the problem of all pairwise comparisons of some function $f(\cdot)$ of G groups (i.e., $f_1 = f_2 = \cdots = f_G$). An example would be all pairwise comparisons of the means of the G groups. The smallest p-value P_1 is compared to the Bonferroni critical value $\alpha_1 = \alpha/R$, where $R = G^*(G-1)/2$. For subsequent tests, however, the family size t_ℓ is determined by the largest number of pairwise hypotheses that can be true simultaneously, given the results of tests $1, \ldots, \ell-1$.

Shaffer actually presented two procedures, commonly referred to as S1 and S2 (e.g., Holland & Copenhaver, 1987, 1988; Rasmussen, 1993). S1 uses the maximum number of hypotheses that could be true at step ℓ, given that $\ell-1$ hypotheses have been rejected. Shaffer gave a simple recursive formula to generate this number.

In the second method, S2, the actual bound t_ℓ is determined conditional on which specific hypotheses have been rejected in steps $1, \ldots, \ell-1$. Consider the case of all pairwise comparisons of seven groups. Having rejected the first hypothesis (e.g., $\mu_3 = \mu_7$), the largest family of hypotheses which can now be true is $\{\mu_3\}\{\mu_1, \mu_2, \mu_4, \mu_5, \mu_6, \mu_7\}$ or $\{\mu_7\}\{\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6\}$. In either case, the largest family size is:

$$\left(\begin{array}{c} 6 \\ 2 \end{array} \right) = 15$$

and this number is used for testing the H_2. If, in testing H_2, the comparison $\mu_3 = \mu_5$ is rejected, then the partition $\{\mu_3\}\{\mu_1, \mu_2, \mu_4, \mu_5, \mu_6, \mu_7\}$ is still logically plausible, $t_3 = 15$, and the third comparison is again performed at $\alpha_3 = \alpha/15$. On the other hand, if the second test rejects the hypothesis $\mu_1 = \mu_4$, then the largest logically consistent partitions are $\{\mu_1, \mu_3\}\{\mu_2, \mu_4, \mu_5, \mu_6, \mu_7\}$, $\{\mu_1, \mu_7\}\{\mu_2, \mu_3, \mu_4, \mu_5, \mu_6\}$, $\{\mu_3, \mu_4\}\{\mu_1, \mu_2, \mu_5, \mu_6, \mu_7\}$ and $\{\mu_3, \mu_7\}\{\mu_1, \mu_2, \mu_4, \mu_5, \mu_6\}$. In each case, the maximum number of null hypotheses that logically can be simultaneously true is:

$$\left(\begin{array}{c} 2 \\ 2 \end{array} \right) + \left(\begin{array}{c} 5 \\ 2 \end{array} \right) = 11.$$

The first method, S1, ignores which hypotheses have been rejected, and so $\alpha_3 = \alpha/15$ for both cases. On the other hand, S2 yields $\alpha_3 = \alpha/15$ in the first case, but $\alpha_3 = \alpha/11$ for the second case. This paper focuses on implementing Shaffer's S2 method.

Both of Shaffer's modifications to the Bonferroni procedure can yield considerable increases in power, but the gain for S2 can be much larger. In the case above

the third test is performed at $\alpha_3 = \alpha/11$ versus $\alpha_3 = \alpha/21$ for the unmodified Bonferroni procedure. This increase in power comes at the price of a great increase in complexity, however. At each step ℓ, the method requires the largest number of null hypotheses which logically may be simultaneously true, given the results of the previous tests $1, 2, \ldots, \ell - 1$. For even moderate number of groups (i.e., 5–10), determining the partition of the groups which corresponds to the largest family size is complex, and for $G > 10$ the prospect is daunting. This complexity has severely limited the use of Shaffer's S2 method, both in applications and in simulation studies.

To get some idea of the complexity involved in using the S2 procedure, consider the data for the pairwise comparisons of 11 cluster analysis methods, presented in Table 1 (this example will be discussed in more detail in Section 5.1). At each stage ℓ, Shaffer's procedure requires the user to determine the partition of the 11 groups into J classes. The partition must be consistent with the previous tests $1, \ldots, \ell - 1$ (i.e., no groups that significantly differ may be placed together) and maximizes the sum $t_\ell = \sum_{j=1}^{J} \binom{n_j}{2}$ where n_j is the number of groups in class j. By the Bonferroni critical value, $H_1 - H_{43}$ have been rejected. Even in this relatively simple case, it is difficult to determine which partitions should be considered. Bear in mind that the application of Shaffer's method requires the solution to this type of problem at each sequential step of testing; several such problems may have to be solved. Clearly, tools are needed to manage the complexity of:

(a) summarizing the results of earlier tests of $H_1 - H_{\ell-1}$; and

(b) determining the partition of the groups yielding the maximum value of t_ℓ.

Such tools will be described in Sections 3 and 4.

1.2.2. Step-up procedures

The earliest step-up procedure (Dunnett & Tamhane, 1992) was proposed for a specific testing situation (comparing treatments with a control) and assumed a multivariate normal distribution. The earliest step-up procedures proposed for more general application are by Hommel (1988) and Hochberg, (1988); an improvement on the latter procedure was derived by Rom (1990). Hochberg and Rom (1995) extended these procedures to allow for logical interrelationships as in Shaffer (1980) described above. FWER control of all of these step-up procedures is based on the validity of Simes inequality (Simes, 1986), which is known to be satisfied for some types of positively dependent test statistics (Sarkar, 1998; Sarkar & Chang, 1997). These methods are known not to control the FWER at the nominal level under all joint distributions, as do the step-down methods. To date, there is no proof that they control the FWER for all pairwise comparisons, although simulation results suggest that they do in many situations of interest (Hochberg & Rom, 1995).

It is relatively easy to show that Holm's (1979) procedure must have more power than the unmodified Bonferroni. Similarly, the Hommel, Hochberg and Shaffer procedures are more powerful than the Holm procedure. Computationally, direct comparison of the procedures has been difficult, although in the specific cases discussed below, Shaffer's procedure is more powerful.

Table 1: Probabilities for the Pairwise Comparisons of 11 Clustering Methods

Comparison	Group 1	Group 2	Sign (G1-G2)	p-value	Significant (Bonferroni)
1	1	3	−	.00000000	Y
2	1	9	−	.00000000	Y
3	1	11	−	.00000000	Y
4	1	8	−	.00000000	Y
5	1	10	−	.00000000	Y
6	1	7	−	.00000000	Y
7	1	6	−	.00000000	Y
8	1	2	−	.00000000	Y
9	1	4	−	.00000000	Y
10	1	5	−	.00000000	Y
11	3	5	+	.00000000	Y
12	5	9	−	.00000000	Y
13	5	8	−	.00000000	Y
14	5	7	−	.00000000	Y
15	5	11	−	.00000000	Y
16	5	10	−	.00000000	Y
17	5	6	−	.00000000	Y
18	2	3	−	.00000000	Y
19	4	5	+	.00000000	Y
20	2	9	−	.00000000	Y
21	2	8	−	.00000000	Y
22	2	11	−	.00000000	Y
23	2	10	−	.00000000	Y
24	2	7	−	.00000000	Y
25	2	5	+	.00000000	Y
26	3	4	+	.00000000	Y
27	2	6	−	.00000000	Y
28	4	9	−	.00000000	Y
29	4	8	−	.00000000	Y
30	3	6	+	.00000000	Y
31	4	11	−	.00000000	Y
32	2	4	−	.00000000	Y
33	3	10	+	.00000000	Y
34	4	7	−	.00000000	Y
35	6	9	−	.00000000	Y
36	4	10	−	.00000000	Y
37	6	8	−	.00000002	Y
38	3	7	+	.00000007	Y
39	6	11	−	.00000103	Y
40	6	7	−	.00000634	Y
41	3	8	+	.00005842	Y
42	3	11	+	.00012791	Y
43	9	10	+	.00020843	Y
44	8	10	+	.00154556	N
45	4	6	−	.00215201	N
46	6	10	−	.00509035	N
47	10	11	−	.00737432	N
48	3	9	+	.00738049	N
49	7	8	−	.05349418	N
50	7	11	−	.06482832	N
51	7	9	−	.08081442	N
52	9	11	+	.16194800	N
53	7	10	+	.17179780	N
54	8	11	+	.20201920	N
55	8	9	+	.88777510	N

2. Implementations of Shaffer's procedure

2.1. Previous implementations

Before proceeding, it is useful to contrast the work presented here to that of Rasmussen (1991, 1993). Rasmussen (1991) presented a computer program for the automatic application of Shaffer's procedure. However, the algorithm (Rasmussen, 1993) makes heavy use of two assumptions: (a) the ordering of the groups has to be consistent with the ordering of the sample means, e.g., $\mu_3 \geq \mu_2 \geq \mu_1$, where the true means are ordered according to the ranks of the sample means; and (b) that the statistical test of the pairwise comparisons exhibit strict transitivity: e.g., μ_2 is significantly greater than μ_1 implies that μ_3 is significantly greater than μ_1 (e.g., Rasmussen, 1993, p. 332). This will be true, for example, if all groups have equal variances. The procedure developed here does not make these assumptions; the order of groups emerges from the analysis. Also, the probability of the tests need not be monotonically related to the group ordering. This might happen if the means of some groups had substantially larger standard errors than did others. Two means that were close together but had small standard errors could yield a smaller p-value than two groups that were further apart but one group had a large standard error.

Shaffer's method has also been implemented by the software package MultComp (Prosoft, 1994). MultComp is an add-on to the SAS statistical package. The current version of the program allows for pairwise comparison of up to 10 groups. From examination of the SAS MACRO code, possible values of t_ℓ are enumerated, as in Shaffer's (1986) article, and therefore MultComp appears to implement Shaffer's S1 method.

A paper by Westfall (1997) considered the problem of using Shaffer's S2 procedure to control for multiple comparisons in the presence of linear contrasts and correlations. Westfall took a geometric approach to evaluating which comparisons are logically possible given the results of previous tests. At step ℓ, the procedure requires the evaluation of $2^{R-\ell} - 1$ subsets, where R is the number of contrasts under consideration. In the present case, $R = \binom{G}{2}$. Clearly, this is only feasible for relatively small problems ($G = 6$ requires over 16,000 evaluations for the first step in testing). However, Westfall's approach is more general in that general linear contrasts may be considered, whereas the present paper considers only the problem of pairwise comparisons. More recently, Westfall, Tobias, Rom Wolfinger, & Hochberg (1999) discuss implementing various multiple comparison procedures using the SAS® system. The book presents software to implement the S2 procedure in SAS. Bretz, Hothorn, & Westfall also present an implementation of the S2 procedure (in the R language).[3]

2.2. The present paper

The present paper first presents a method of representing the results of the comparisons made. Next, several properties of an optimal partition of the groups (which maximizes t_ℓ) are derived; note that the optimal partition need not be unique (see Section 3.3 below). The use of these results in determining potential candidates for the partition of groups that corresponds to the largest family size is described. Next, two heuristics are described which vastly limit the search space, greatly enhancing the efficiency of the determination of t_ℓ. Finally, an efficient algorithmic method

[3] One of the reviewers pointed out that "... these software not only provide Shaffer S2, but uniform improvements upon it under the usual linear model assumptions by utilizing the multivariate t-distribution, rather than Bonferroni's method."

of determining t_ℓ is outlined. These methods are illustrated by application to two data sets: comparisons of the order of 11 clustering methods (Donoghue, 1995); and comparison of the means of 44 states from the 1994 National Assessment of Educational Progress Trial State Assessment of Reading at Grade 4 (Williams, Reese, Campbell, Mazzeo, & Phillips, 1995).

3. Theoretical results

3.1. Terminology

Let $\mathbf{X}$ be a collection of G groups (i.e., distributions) $X_1, \ldots, X_g, \ldots, X_G$. We wish to test all pairwise comparisons of some function $f(\cdot)$ of the G distributions. Let $\mathbf{P}$ be the vector of p-values for the pairwise comparisons of the groups, and let $\mathbf{H}$ be the associated hypotheses permuted in the same order as $\mathbf{P}$. The dimension of both $\mathbf{P}$ and $\mathbf{H}$ is $R = G^*(G-1)/2$.

Assume that the test statistic $T_{gg'} = T(f(X_g), f(X_{g'}))$ yields both a test of the significance of the difference between groups g and g' and an indication of the direction, i.e., $T_{gg'} > 0 \Rightarrow f(X_g) > f(X_{g'})$. The most common example is the t-test, although other descriptive statistics (e.g., Cliff's (1993) d-statistic) may also suffice. The requirement of order is not necessary for the present method, but does facilitate the discussion.

3.1.1. Adjacency matrices

To represent the results of comparisons, we make use of ideas from Cliff:[4] (a) adjacency matrices (Cliff, 1975); and (b) dominance matrices (Cliff, 1993). An adjacency matrix

> "... is simply a matrix for which the existence or nonexistence of the connection or relation is recorded. For a set X of n elements and a relation R, the n x n adjacency matrix $\mathbf{A}$ has $a_{ij} = 1$ if i R j and $a_{ij} = 0$ otherwise." (Cliff, 1975, p. 290).

Here, the relation will be "significantly greater than," and $a_{gg'} = 1$ if $P(T > |T_{gg'}|) \geq \alpha'/2$ and $f(X_g) > f(X_{g'})$ (although the relation is stated in terms of a directional hypothesis, significance will typically be determined using two-tailed tests). Adjacency matrices have several other properties; in this context we will only use them as a notational device. Cliff notes that a simple order of the elements has a unique representation. If $x > y > z$, then:

$$\mathbf{A} = \begin{array}{c} \\ x \\ y \\ z \end{array} \begin{array}{c} \begin{array}{ccc} x & y & z \end{array} \\ \left[\begin{array}{ccc} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right] \end{array} .$$

To illustrate the usefulness of the adjacency matrix, Table 2 presents an adjacency matrix summarizing the results of the tests in Table 1.

Although it is not necessary in order to compute Shaffer's procedure, the relations in $\mathbf{A}$ are simpler to comprehend if the rows and columns of $\mathbf{A}$ are sorted to approximate upper triangular form. If available, this can be accomplished by sorting based on some appropriate statistic, such as the group means. Alternatively,

[4]Both of these ideas have appeared in other contexts. However, their use in this paper will be consistent with the treatment of the ideas in Cliff.

Table 2: Adjacency Matrices for Tests in Table 1

	A	B	C	D	E	F	G	H	I	J	K
A	–	0	0	0	0	0	0	0	0	0	0
B	1	–	0	0	1	0	0	0	0	0	0
C	1	1	–	1	1	1	1	1	0	1	1
D	1	1	0	–	1	0	0	0	0	0	0
E	1	0	0	0	–	0	0	0	0	0	0
F	1	1	0	0	1	–	0	0	0	0	0
G	1	1	0	1	1	1	–	0	0	0	0
H	1	1	0	1	1	1	1	–	0	0	0
I	1	1	0	1	1	1	1	0	–	1	0
J	1	1	0	1	1	1	0	0	0	–	0
K	1	1	0	1	1	1	1	0	0	0	–

Original

	C	I	G	H	K	J	F	D	B	E	A
C	–	0	1	1	1	1	1	1	1	1	1
I	0	–	0	0	0	1	1	1	1	1	1
G	0	0	–	0	0	0	1	1	1	1	1
H	0	0	0	–	0	0	1	1	1	1	1
K	0	0	0	0	–	0	1	1	1	1	1
J	0	0	0	0	0	–	0	1	1	1	1
F	0	0	0	0	0	0	–	0	1	1	1
D	0	0	0	0	0	0	0	–	1	1	1
B	0	0	0	0	0	0	0	0	–	1	1
E	0	0	0	0	0	0	0	0	0	–	1
A	0	0	0	0	0	0	0	0	0	0	–

Sorted by Potency Vector **p**

Undifferentiated Classes (Bonferroni)

$$
\begin{aligned}
\mathrm{E}_1 &= \{\mathrm{C}, \mathrm{I}\} \\
\mathrm{E}_2 &= \{\mathrm{I}, \mathrm{G}, \mathrm{H}, \mathrm{K}\} \\
\mathrm{E}_3 &= \{\mathrm{G}, \mathrm{H}, \mathrm{K}, \mathrm{J}\} \\
\mathrm{E}_4 &= \{\mathrm{J}, \mathrm{F}\} \\
\mathrm{E}_5 &= \{\mathrm{F}, \mathrm{D}\} \\
\mathrm{E}_6 &= \{\mathrm{B}\} \\
\mathrm{E}_7 &= \{\mathrm{E}\} \\
\mathrm{E}_8 &= \{\mathrm{A}\} \\
\\
\Pi &= \mathrm{s}(\mathrm{E}_3, \mathrm{E}_1, \mathrm{E}_5, \mathrm{E}_6, \mathrm{E}_7, \mathrm{E}_8)\text{: } t_\ell = 8
\end{aligned}
$$

the matrix may be ordered based on dominance relations. Dominance relations are defined as:

$$
d_{gg'} = \begin{cases} 1, f(x_g) > f(x_{g'}) \\ 0, f(x_g) = f(x_{g'}) \\ -1, f(x_g) < f(x_{g'}) \end{cases}
$$

The matrix of dominance relations is $\mathbf{D} = \mathbf{A} - \mathbf{A'}$.[5] Define the potency vector $\mathbf{p} = \mathbf{D1}$. Experience indicates that sorting based on $\mathbf{p}$ will tend to maximize the agreement with an upper triangular form of $\mathbf{A}$, while minimizing any discrepancies in ordering, although I am not aware of any rigorous proof. The adjacency matrix on the right hand side of Table 2 is sorted based on $\mathbf{p}$. In this form, the patterns of groups that do and do not differ from one another are much simpler to comprehend.

3.1.2. Undifferentiated classes

A fundamental idea for much of the following is a set of groups such that none of the groups differs significantly from any other (e.g., for means, we cannot reject $\mu_1 = \cdots = \mu_b = \cdots = \mu_k$). At an arbitrary step ℓ, the term "undifferentiated class" will be used to denote the largest set of groups that do not (thus far in the testing) differ

[5] We note in passing that a more visual form of the information in the dominance matrix was used to report results for individual states in the 1990, 1992, and 1994 NAEP Trial State Assessments.

significantly from each other. The term largest implies that for any group g' which is not in undifferentiated class, $T_{gg'}$ is significant for at least one element g in the undifferentiated class. The adjacency matrix represents an undifferentiated class as a square submatrix (symmetric about the main diagonal) of $\mathbf{A}$ whose elements are all zero, and is obtained by extracting the proper rows and corresponding columns of $\mathbf{A}$. The bottom section of Table 2 presents the undifferentiated classes for the tests presented in Table 1.

3.2. Notation

At step ℓ in the sequential testing, consider the collection of Q (nonempty) undifferentiated classes $E_q = \{X_g, g = 1, \ldots, n_q\}. E_q \neq \emptyset \forall q, E_q \cap E_{q'}$ need not be empty, and $\bigcup_{q=1}^{Q} E_q = X$. Let $\Pi = \{\pi_1, \ldots, \pi_J\}$ be a partition of X, e.g., $\pi_j \subseteq X \forall j, \pi_j \cap \pi_{j'} = \emptyset \forall j, j'$ and $\bigcup_{j=1}^{J} \pi_j = X$. The notation Z^{C} will indicate the complement of set Z on X.

Let $N_q = N(Z_q)$, the number of elements in set Z_q. Similarly, define $n_j = N(\pi_j)$ for a given element j in a partition Π (use of the lower case n will emphasize that the argument is part of a partition). The family size τ for a given partition is:

$$\tau = \sum_{j=1}^{J} \binom{n_j}{2}$$

with the notational convenience:

$$\binom{1}{2} = 0.$$

An "optimal partition" at step ℓ, is a partition $\Pi(\mathbf{X})$ that yields t_ℓ the maximum value of τ. Note that:

$$\tau = \sum_{j=1}^{J} \frac{n_j^2 - n_j}{2} = \sum_{j=1}^{J} \frac{n_j^2}{2} - \sum_{j=1}^{J} \frac{n_j}{2} = \frac{1}{2} \sum_{j=1}^{J} n_j^2 - \frac{G}{2}.$$

Thus, maximizing

$$s = \sum_{j=1}^{J} n_j^2 = 2\tau + G$$

is equivalent to maximizing τ. The remainder of the paper will focus on obtaining a partition Π which yields $\sigma = s(\Pi)$ the maximal value of $s(.)$ at step ℓ in the sequential testing.

3.3. Properties of the optimal partition

This section presents four theorems about the structure of an optimal partition. These theorems will be used in Section 3.4 to develop a procedure to compute t_ℓ. First, some obvious properties of the optimal partition are enumerated. These properties will prove useful below.

Property 1. Each element of the optimal partition Π is a subset of (at least) one of the undifferentiated classes: for each $j, \pi_j \subseteq E_q$ for some q. This property simply corresponds to the formulation of Shaffer's procedure. If π_j is not a subset of any undifferentiated class, then the partition is not consistent with the results of the tests at stages $1, \ldots, \ell - 1$.

Property 2. No subpartition of a partition element π_j into two subpartitions can yield a higher value of $s(.)$ than is obtained by using the intact element. The proof of this property is by contradiction. Let π_j be the element in question, and $n_j = N(\pi_j)$. Let the subpartitions be π_{jA} and π_{jB}, with $\pi_{jA} \subset \pi_j, \pi_{jB} \subset \pi_j, \pi_{jA} \cup \pi_{jB} = \pi_j$, and $\pi_{jA} \cap \pi_{jB} = \emptyset$. Let $n(\pi_{jA}) = kn_j$ and $n(\pi_{jB}) = (1-k)n_j$, $0 < k < 1$. Assume:

$$s(\pi_j) < s(\pi_{jA}) + s(\pi_{jB}).$$

Then

$$\begin{aligned} n_j^2 &< k^2 n_j^2 + (1-k)^2 n_j^2 \\ 0 &< (2k^2 - 2k) n_j^2 \\ 0 &< (k-1)k n_j^2. \end{aligned}$$

The right hand side must be negative, yielding a contradiction and so proving Property 2.

Property 3. No subpartition of an element of a partition yields a larger value of $s(\cdot)$ than is obtained by using the intact element. This property follows immediately from repeated application of Property 2.

Property 4. There is a one-to-one relationship between the elements of the optimal partition and the undifferentiated classes: $\pi_q \subseteq E_q \forall q$ (note that π_q may be empty for some undifferentiated classes E_q). This property follows directly from Properties 1 and 3.

Theorem 1. *Let the set* $\boldsymbol{X}$ *consist of two classes,* E_i *and* E_j*. Choose* i *and* j *such that* $N(E_i) \geq N(E_j)$*. Then the partition* $\{\pi_1, \pi_2\}$*, where* $\pi_1 = E_i, \pi_2 = E_j \cap E_i^C$*, is optimal.*

Proof. The proof is by contradiction. If Theorem 1 is false, then there exists some partition of $(E_i \cup E_j)$ for which assigning some (or all) of the intersection to E_j yields a higher value of $s(.)$ than s $(E_i, E_j \cap E_i^C)$. Let:

$$\begin{aligned} m_{ij} &= N(E_i \cap E_j) \\ m_i &= N\big(E_i \cap E_j^C\big) = N_i - m_{ij} \\ m_j &= N\big(E_j \cap E_i^C\big) = N_j - m_{ij} \end{aligned}$$

Assume Theorem 1 is false. Then for some value k, such that $0 \leq k < 1$:

$$\begin{aligned} & s\big(E_i, E_j \cap E_i^C\big) < s(\pi_1, \pi_2) \\ & (m_i + m_{ij})^2 + m_j^2 < (m_i + k m_{ij})^2 + \big(m_j + (1-k)m_{ij}\big)^2 \\ & m_i^2 + 2m_i m_{ij} + m_{ij}^2 + m_j^2 \\ & \qquad < m_i^2 + 2k m_i m_{ij} + k^2 m_{ij}^2 + m_j^2 + 2(1-k)m_j m_{ij} + (1-k)^2 m_{ij}^2 \\ & 0 < 2(1-k)m_{ij}(m_j - m_i) + (1-k)^2 - \big(k^2 - 1\big)m_{ij}^2 \\ & 0 < 2(1-k)m_{ij}(m_j - m_i) + 2k(k-1)m_{ij}^2 \end{aligned}$$

The first term on the right hand side is nonpositive by hypothesis: $N(E_i) \geq N(E_j) \Rightarrow m_i \geq m_j$. The second term is nonpositive because $0 \leq k < 1$. This yields the contradiction that the right side is greater than zero, and hence Theorem 1 is true. □

Theorem 2 (Converse). *Again, let* $\boldsymbol{X}$ *consist of two classes,* E_i *and* $E_j, N(E_i) > N(E_j)$*. Let* $\{\pi_1, \pi_2\}$ *be an optimal partition on* $\boldsymbol{X}$*. Then, either: (Case* 1*)* $\pi_1 \subseteq E_i, \pi_2 \subseteq E_j$*, and* $\pi_1 = E_i, \pi_2 = E_j \cap E_i^C$*; or (Case* 2*)* $\pi_1 \subseteq E_j, \pi_2 \subseteq E_i$*, and* $\pi_1 = E_j \cap E_j^C, \pi_2 = E_i$*.*

Proof. (Case 1) Let $\pi_1 \subseteq E_i$ and define m_i, m_j, and m_{ij} as in the Proof of Theorem 1 above. If $\pi_1 \subseteq E_i$ then there exists $k(0 \leq k \leq 1)$ such that $n(\pi_1) = n_1 = m_1 + km_{ij}$ and $n(\pi_2) = n_2 = m_2 + (1-k)m_{ij}$.

$$\begin{aligned} s(\pi_1, \pi_2) &= n_1^2 + n_2^2 \\ &= m_i^2 + 2km_i m_{ij} + k^2 m_{ij}^2 + m_j^2 + 2(1-k)km_j m_{ij} + (1-k)^2 m_{ij}^2 \end{aligned}$$

and

$$s\big(E_i, E_j \cap E_i^C\big) = m_i^2 + 2m_i m_{ij} + m_{ij}^2 + m_j^2$$

By definition: $s(E_i, E_j \cap E_i^C) \leq s(\pi_1, \pi_2)$

$$\begin{aligned} m_i^2 + 2m_i m_{ij} + m_{ij}^2 + m_j^2 &\leq m_i^2 + 2km_i m_{ij} + k^2 m_{ij}^2 + m_j^2 + 2(1-k)m_j m_{ij} \\ &\quad + (1-k)^2 m_{ij}^2 \\ 0 &\leq 2m_{ij}(k-1)(m_i - m_j) + 2m_{ij}^2 k(k-1) \end{aligned}$$

The value $0 \leq k < 1$ leads to a contradiction, and equality holds only if $k = 1$. Thus, $\pi_1 \subseteq E_i$ implies $\pi_1 = E_i$

(Case 2) Let $\pi_1 \subseteq E_j$. Then, as above, $n(\pi_1) = n_1 = m_1 + km_{ij}$ and $n(\pi_2) = n_2 = m_2 + (1-k)m_{ij}$, and $s(\pi_1, \pi_2) \geq s(E_i, E_j \cap E_i^C)$. As before:

$$\begin{aligned} m_i^2 + 2m_i m_{ij} + m_{ij}^2 + m_j^2 &\leq m_j^2 + 2km_j m_{ij} + k^2 m_{ij}^2 + m_i^2 + 2(1-k)m_i m_{ij} \\ &\quad + (1-k)^2 m_{ij}^2 \\ 0 &\leq 2m_{ij}k(m_j - m_i) + 2m_{ij}^2 k(k-1) \end{aligned}$$

and $0 < k \leq 1$ leads to a contradiction. For $k = 0$, equality holds and implies $\pi_2 = E_i$ and $\pi_1 = E_j \cap E_i^C$ which is the hypothesized partition. □

Property 5 (Local optimality). For an optimal partition with elements $\{\pi_1, \pi_2, \ldots, \pi_j, \ldots, \pi_Q\}, \sigma = n(\pi_1)^2 + n(\pi_2)^2 + \cdots + n(\pi_j)^2 + \cdots + n(\pi_Q)^2$. Define the partial sum $\sigma_{ij} = s(\Pi) - n(\pi_i)^2 - n(\pi_j)^2$. Thus, $\sigma(\mathbf{X}) = \sigma_{ij} + s(\pi_{(i)}, \pi_{(j)})$. The optimality of $\sigma = s(\Pi)$ implies that $s(\pi_{(i)}, \pi_{(j)}) = \sigma(\pi_i, \pi_j)$ for all pairs i, j. Indeed, for any disjoint subset of $\mathbf{X}$ consisting of the union of elements of an optimal partition, the value of $s(.)$ obtained from the partition must be optimal for that subset. This property will be termed local optimality, and in the specific case of two elements, pairwise local optimality.

Property 6 (Nonuniqueness of the optimal partition). Theorem 2 points out that the optimal partition is not uniquely determined. If Π is an optimal partition of $\mathbf{X}$, then any permutation of the elements $\{\pi_1, \pi_2, \ldots \pi_Q\}$ yields an identical value of $s(.)$, and is also an optimal partition. The order of the elements is not the only source of ambiguity. For $\mathbf{X}$ consisting of two classes E_i and $E_j, N(E_i) = N(E_j), s(E_i, E_j \cap E_i^C) = s(E_j, E_i \cap E_j^C)$ and both are optimal partitions of $\mathbf{X}$.

Because of Property 6, many results about maximizing $s(\mathbf{X})$ must be stated in terms of the existence of an optimal partition with specific properties. All optimal partitions need not possess those properties. Thus, Theorems 3 and 4 below are stated in terms of the existence of an optimal partition with specific properties, rather than describing the properties of optimal partitions.

Theorem 3. *Consider the set* $\boldsymbol{X}$. *At step* ℓ *in the sequential testing, undifferentiated classes* $\boldsymbol{E} = \{E_q, q = 1, Q\}$ *are obtained, and yield a maximum value of* $s = \sigma$. *There exists an optimal partition* $\Pi(\mathbf{X}) = \{\pi_1, \pi_2, \ldots, \pi_q, \ldots, \pi_Q\}$ *which, for appropriate indexing of the undifferentiated classes, has the form:* $\pi_q = E_q \bigcap (\bigcup_{v=1}^{q-1} E_v)^C$, *and the optimal partition is:*

$$\left\{ E_1, E_2 \bigcap E_1^C, E_3 \bigcap \left(E_1 \bigcup E_2 \right)^C, \ldots, E_Q \bigcap \left(\bigcup_{q=1}^{Q-1} E_q \right)^C \right\}.$$

Proof. Let the vector $\mathbf{n_{(i)}}$ be the order statistic of $n(\Pi)$, with $\pi_{(i)}$ representing the corresponding permutations of the elements of Π; that is $n_{(1)} = N(\pi_{(1)}) \geq N(\pi_{(2)}) \geq \cdots \geq N(\pi_{(Q)}) = n_{(Q)}$. By Property 4, there is a one-to-one correspondence between each element π_q and E_q. Therefore, let $E_{(i)}$ represent the permutation of the undifferentiated classes which corresponds to $\pi_{(i)}$: i.e., $\pi_{(i)} \subseteq E_{(i)}$. Finally, let $\eta_{ij} = E_{(i)} \cap (\pi_{(i)} \cup \pi_{(j)})$, and define η_{ji} analogously. The remainder of the proof consists of establishing four connected results, denoted Result 3A–3D.

Result 3A. For $i < j, N(\eta_{ij}) \geq N(\eta_{ji})$. If this were not true, Theorem 1 states that $s(\eta_{ji}, \eta_{ij} \cap \eta_{ji}^C)$ would be larger than $s(\eta_{ij}, \eta_{ji} \cap \eta_{ij}^C)$, which is contrary to the local pairwise optimality of $\pi_{(i)}$ and $\pi_{(j)}$.

Result 3B. For $i < j, \sigma(\eta_{ij}, \eta_{ji}) = s(\eta_{ij}, \eta_{ji} \cap \eta_{ij}^C)$ by Result 3A and Theorem 1, and so $\pi_{(i)} = \eta_{ij}$. This gives:

$$\begin{aligned} \pi_{(i)} &= \eta_{ij} \\ &= E_{(i)} \cap \left(\pi_{(i)} \cup \pi_{(j)} \right) \\ &= \left(E_{(i)} \cap \pi_{(i)} \right) \cup \left(E_{(i)} \cap \pi_{(j)} \right) \\ &= \pi_{(i)} \cup \left(E_{(i)} \cap \pi_{(j)} \right) \end{aligned}$$

which implies $E_{(i)} \cap \pi_{(j)} = \emptyset$ for all $j > i$.

Result 3C. Result 3B implies that $\eta_{1j} = \pi_{(1)}$ and that $E_{(1)} \cap \pi_{(j)} = \emptyset \forall j > 1$. Because Π is a complete partition of $\mathbf{X}$ (i.e., $\bigcup_{q=1}^{Q} \pi_q = X$), $\pi_{(1)} = E_{(1)}$.

Result 3D. Result 3C shows that $\pi_{(2)}$ cannot overlap $E_{(1)}$, and Result 3B indicates that $E_{(2)} \cap \pi_{(j)} = \emptyset$ for all $j > 2$. Therefore, $\pi_{(2)} = E_{(2)} \cap E_{(1)}^C$. Similarly, recursive application of Result 3C and Result 3B thus yields the full form of

$$\Pi = \left\{ E_1, E_2 \bigcap E_1^C, E_3 \bigcap \left(E_1 \bigcup E_2 \right)^C, \ldots, E_Q \bigcap \left(\bigcup_{q=1}^{Q-1} E_q \right)^C \right\},$$

establishing Theorem 3.

□

Note that, although $n(\pi_{(i)}) \geq n(\pi_{(j)})$ for i ¡ j, $N(E_{(i)})$ need not be less than $N(E_{(j)})$. For example, consider the adjacency matrix shown in Table 3. Note that, although $N_2 > N_1, N_3$, the partition $\{1, 2, 3, 4\}$ $\{5, 6, 7, 8\}$ yields a larger value, $s = 32$, than does the partition $\{1, 2, \}$ $\{3, 4, 5, 6, 7\}$ $\{8\}$, which has $s = 30$.

Table 3: Adjacency Matrix for Hypothetical Example of 8 Groups

	1	2	3	4	5	6	7	8
1	–	0	0	0	1	1	1	1
2	0	–	0	0	1	1	1	1
3	0	0	–	0	0	0	0	1
4	0	0	0	–	0	0	0	1
5	0	0	0	0	–	0	0	0
6	0	0	0	0	0	–	0	0
7	0	0	0	0	0	0	–	0
8	0	0	0	0	0	0	0	–

Undifferentiated Classes

$\mathrm{E}_1 = \{1, 2, 3, 4\}$
$\mathrm{E}_2 = \{3, 4, 5, 6, 7\}$
$\mathrm{E}_3 = \{5, 6, 7, 8\}$

$s\big(E_1, E_3, E_2 \cap (E_1 \cup E_3)^C\big) = 4^2 + 4^2 + 0^2 = 16 + 16 + 0 = 32$
$s\big(E_2, E_1 \cap E_2^C, E_3 \cap E_2^C\big) = 5^2 + 2^2 + 1^2 = 25 + 4 + 1 = 30$

Theorem 4. *There exists an optimal partition* Π *with the structure given in Theorem 3 for which the* $n_{(j)} = n(\pi_{(j)})$ *are sorted from largest to smallest, i.e.,* $n_{(i)} \geq n_{(j)} \forall i < j$.

Proof. The proof follows directly from repeated application of Theorem 1 to the structure of Theorem 3: $n_{(1)} \geq n_{(2)}, n_{(2)} \geq n_{(3)}$, etc. Combining these relations gives $n_{(1)} \geq n_{(2)} \geq \cdots \geq n_{(J)}$. □

Theorems 1–4 are sufficient to yield a workable procedure to implement Shaffer's S2 procedure for a large (~ 50) number of groups. Stronger results concerning the optimal partition have been obtained, and doubtless, still other results could be derived. However, they are not necessary for the present purpose, and so will not be discussed here.

4. Implementation of Shaffer's procedure

4.1. Brute force enumeration

Theorem 3 gives us an algorithmic (albeit inefficient) method of deriving an optimal partition of $\mathbf{X}$. We simply compute s for the $Q!$ partitions of the form given by Theorem 3, and select the maximum value. However, this is needlessly intensive, for often the first $J (J \ll Q)$ elements will from a full partition of the groups, and thus there is no need to compute the $(Q - J)!$ arrangements of classes with $n_j = 0$ (nor indeed need elements with $n_j = 1$ be considered, for their contribution to s must always be 1.) For example, consider the comparison of 44 state means taken from the 1994 National Assessment of Educational Progress Trial State Assessment (NAEP TSA) in reading. At one step in the sequential testing, omitting consideration of elements with $n_j = 0$ reduced the number of comparisons from 1.3 trillion to approximately 17.7 million, a change from 185 years of CPU time on a SUN SPARCstation to approximately 24 hours.

4.2. Heuristic 1: Forward pruning of the search

When used in conjunction with the pairwise local optimality property, Theorem 1 is very powerful in allowing us to perform forward pruning of the tree as we search for an optimal partition. Consider a partial partition which begins with groups $1, 2, \ldots, j-1$ and for which $n_{j-1} < n_j$. We can delete (without enumeration or calculating the value of s) the $(Q-j)!$ permutations of the remaining groups. These permutations can be eliminated because the permutation $1, \ldots, j, j-1, \ldots, Q$ will have a value of $s(.)$ which is at least as large as the permutation $1, \ldots, j-1, j, \ldots, Q$. Consider the first two elements of the permutation. There are $G^*(G-1)$ permutations, but one-half of them can be rejected immediately, because they involve $n_1 \leq n_2$. By Theorem 1, these partitions cannot yield a larger value of $s(.)$, and so need not be examined. Application of Theorem 1 to each adjacent pair of groups in the search yields a similar halving of the search space. Note that Theorem 1 implies that the order of two groups of equal size is irrelevant. They are therefore ordered by their indices, and only those partitions in which the lower indexed member of a tied pair occurs first need be considered.

4.3. Heuristic 2: A killer heuristic

Theorem 4 is useful in the form of a killer heuristic. Using only Theorems 1 and 3, it is possible to generate large searches that are obviously suboptimal. In the case of the comparison of 44 state means in the NAEP TSA reading assessment, for example, searches that begin with $n_1 = 3$ or 4 clearly yield smaller total values of $s(.)$ than those beginning with $n_1 = 18$. However, using only Theorem 1, such permutations would have to be pursued to great depth before they could be shown to be inferior. By Theorem 4 however, n_j can be used to place an upper bound on the values of $n_{j+1}, \ldots, n_Q$.

Theorem 4 can be used compute an upper bound on the value of the full partition computed from the current partial partition:

$$s_{\max -1} = \sum_{v=1}^{j} n_v^2 + n_j^2 + \left(G - \sum_{v=1}^{j} n_v^2 + K \cdot n_j^2 \right)^2$$

where K is the greatest integer not greater than $\frac{G - \sum_{v=1}^{j} n_v}{n_j}$. If $s_{\max}$ is less than or equal to the lower bound on σ, the partial permutation may be discarded without further searching. Actually, $s_{\max -1}$ is too large. A better (smaller) upper bound for the partial partition is:

$$s_{\max -2} = \sum_{k=1}^{j-1} n_k^2 + n_j^2 + \sum_{v=j+1}^{V} \min\left(n_j^2, n_v^2\right),$$

where V is the integer such that $\sum_{v=1}^{V} n_v < G \leq \sum_{v=1}^{V+1} n_v$. In this equation, n_v is based on removing the overlap with groups $1-j$; it is not a true $n(\pi_v)$ and so $s_{\max -2}$ is an upper bound.

All that remains to make Theorem 4 useful is $\hat{\sigma}_{\min}$, a lower bound on σ. The value of $s(.)$ computed for any partition is a lower bound on σ. However, the larger the value of $\hat{\sigma}_{\min}$ the more useful it will be in limiting the search. A good compromise between simplicity and yielding the largest possible value is:

$$\hat{\sigma}_{\min} = \sum_{j=1}^{J} n \max_{L \neq 1 \ldots j} \left(E_L \bigcap \left(\bigcup_{i=1}^{j} E_i \right)^C \right).$$

It is possible (due to ties) that some permutations will have to be evaluated in determining $\hat{\sigma}_{\min}$. It is not necessary to explicitly maximize $\hat{\sigma}_{\min}$; its only purpose is to serve as a lower bound on σ with which to compare $s_{\max -2}$. In practice, however, it has seemed wise to evaluate it fully, because Theorem 4 is tested frequently and responsible for much of the reduction in computation.

The combined used of Theorem 1 and Theorem 4 results in huge reductions in the search to find σ. Typical analyses on the 40-45 jurisdictions in the samples from different years of NAEP TSA data require examination of 10-30 permutations to determine $\hat{\sigma}_{\min}$. The full search has then required 0-4 evaluations of $s(.)$ for full permutations. Run time on a 133 MHz Pentium machine is typically 5-15 seconds for a complete analysis of the pairwise comparisons of 45 jurisdictions in the NAEP TSA samples.

4.4. A simplified algorithm

Using Theorems 1–4, it is relatively straightforward to construct an algorithm to compute the value of σ for the pairwise comparisons of a dataset consisting of G groups. This section outlines the general approach. A companion paper will describe the details of the complete algorithm programmed to implement Shaffer's procedure.

1) Compute initial $t_\ell = G*(G-1)/2$ (the Bonferroni or Holm critical value may be used)
2) Perform tests using t_ℓ
3) Record results in adjacency matrix **A**
4) If no new differences then STOP
5) Else: Extract undifferentiated classes from **A** (represented by square submatrices of zeroes which are symmetric about the main diagonal)
6) Form $\hat{\sigma}_{\min}$
   A. Initialize $\hat{\sigma}_{\min}$ to zero
   B. Compute group sizes
   C. Find largest size group, k.
   D. If $n_k = 0$ then Record permutation and Exit.
   E. Else: $\hat{\sigma}_{\min} = \hat{\sigma}_{\min} + n_k^2$
   F. Remove overlapping elements of group k from all other groups
   G. Go to B
7) Compute σ
   A. Initialize: $s_{\max} = \hat{\sigma}_{\min}$; $q = 1$;
   B. Compute permutation
      i) Set $j = 1$
      ii) Select element j of the partition to be class $E_q : \pi_j = E_q \cap (\bigcup_{k=1}^{j-1} E_k)^C$, $s = s + n_q^2$, $n_j = n_q$, $n_q = 0$
      iii) Remove overlapping groups from other classes k and recompute class sizes n_k for all $k \neq j$

iv) If $\max(n_k) = 0$ for all $k \neq j$ then this is a complete partition. Go to C.

v) Else look for next element $k : k = 0$

a) $k = k + 1$

b) Test 1: Is k valid
If $k > G$ then no optimal permutation beginning $\pi_1, \ldots, \pi_j$. Go to D
If E_k is already in use then Go to B(v)(a)

c) Test 2–Theorem 2 (Killer Heuristic 1)
If $n_k > n_j$ then Go to B(v)(a)

d) If $n_k = n_j$ and $k > j$ then Go to B(v)(a)[6]

e) Test 3–Theorem 3 (Killer Heuristic 2)
Test 3A:
Compute K
If $s + K * n_k^2 < \hat{\sigma}_{\min}$ then no optimal permutation begins $\pi_1, \ldots, \pi_j$. Go to D
Test 3B:
If $s + \sum_{v \neq k}^{Q} \min(n_k, n_v)^2 < \hat{\sigma}_{\min}$ then no optimal permutation begins $\pi_1, \ldots, \pi_j$. Go to D

vi) Accept k as next element of partition: $q = k, j = j + 1$; Go to B(iii).

C) If $s > s_{\max}$ then $s_{\max} = s$; Record permutation

D) Done with $\pi_1, \ldots, \pi_j$. Find next j.

i) Remove element $\pi_j : q = \pi_j, j = j - 1$

ii) $q = q + 1$

iii) Test 1: Is q valid
a) If E_q is already in use then Go to D(ii)
b) If $q > G$ then no new valid permutation beginning $\pi_1, \ldots, \pi_j$.
Test 2: Is j valid?
If $j > 0$ then Go to D(i)
If $j = 0$ and $q > G$ then all permutations have been examined.
Go to 8.

iv) The partial partition is valid. Go to B(iii)

8) Done. Set $\sigma = s_{\max}$. Print permutation.

9) Compute $t_\ell = \frac{\sigma}{2} - \frac{G}{2}$. Go to 2

5. Examples

A computer program has been written based on the logic above. This section presents the results of applying the program to selected datasets.

[6]This test prohibits the unnecessary evaluation of permutations of pairs of undifferentiated classes of equal size.

5.1. Example 1

Example 1 is taken from a simulation study of the ability of several cluster analysis procedures to recover known subgroups (Donoghue, 1995). The "groups" are the cluster analysis methods, and the outcome variable cluster recovery as measured by the Hubert and Arabie (1985) modification of the Rand (1971) index. Pairwise ordinal comparisons of the methods were made using Cliff's (1993) ordinal method for paired comparisons, computed using the PAIRDEL program (Cliff, 1992). Each comparison yields a t-test for the pair. Unlike the situation considered by Rasmussen (1993), there is no statistic upon which the methods may be ordered (indeed, determining such an order while maintaining FWER was the chief goal of the analysis). Table 1 presents the probabilities for each comparison under the null hypothesis that the methods are not ordered. Table 2 shows the initial adjacency matrix (based on the Bonferroni critical value) associated with these t-tests; Table 4 shows the final adjacency matrix. The rows and columns have been permuted according to the potency vector **p** (see Section 3.1.1), and minus signs have been substituted into the diagonal to aid in the visual presentation. The "1" symbol indicates those comparisons that were found to be significant using the Bonferroni correction; the "X" entries are comparisons which were found to be significant by Shaffer's procedure but not by the Bonferroni procedure.

Table 4: Adjacency Matrix for Example 1: Comparison of 11 Cluster Analysis Methods

	C	H	I	K	G	J	F	D	B	E	E
C	−	X	1	1	1	1	1	1	1	1	1
H	0	−	0	0	0	1	1	1	1	1	1
I	0	0	−	0	0	X	1	1	1	1	1
K	0	0	0	−	0	X	1	1	1	1	1
G	0	0	0	0	−	0	1	1	1	1	1
J	0	0	0	0	0	−	X	1	1	1	1
F	0	0	0	0	0	0	−	X	1	1	1
D	0	0	0	0	0	0	0	−	1	1	1
B	0	0	0	0	0	0	0	0	−	1	1
E	0	0	0	0	0	0	0	0	0	−	1
A	0	0	0	0	0	0	0	0	0	0	−

Undifferentiated Classes (Final)

$$\mathrm{E}_1 = \{\mathrm{H}, \mathrm{I}, \mathrm{K}, \mathrm{G}\} \quad \mathrm{E}_3 = \{\mathrm{A}\} \quad \mathrm{E}_5 = \{\mathrm{C}\} \quad \mathrm{E}_7 = \{\mathrm{E}\}$$
$$\mathrm{E}_2 = \{\mathrm{G}, \mathrm{J}\} \quad \mathrm{E}_4 = \{\mathrm{B}\} \quad \mathrm{E}_6 = \{\mathrm{D}\} \quad \mathrm{E}_8 = \{\mathrm{F}\}$$

$$\Pi = \mathrm{s}\big(E_1, E_2 \cap E_1^C, E_3, E_4, E_5, E_6, E_7, E_8\big) : t_\ell = 6$$

Procedure	P_{critical}	Number of hypotheses not rejected
Bonferroni	.0009	12
Holm	.0045	10
Hochberg	.0045	10
Hommel	.0063	9
Shaffer	.0083	7

The bottom section of Table 2 presents the undifferentiated classes for the Bonferroni comparisons. In classes E_4 and E_5, groups J and F, and F and D form undifferentiated classes of size 2 each. Note that, because $a_{68} = 1$, the three elements J, F and D do *not* form an undifferentiated class of size 3, but only two separate classes of size 2. Further, it is not possible for both of these relations to be true simultaneously; either $f_J = f_F$ or $f_F = f_D$. If both were true, this would imply $f_J = f_D$, which has been rejected. Because both cannot be true, only one needs to be counted in determining the family size for the next comparison. It is through noting such relations that Shaffer's procedure may obtain substantial increases in power over Holm's procedure.

Table 4 presents the final undifferentiated classes from Shaffer's procedure. These may be useful in reporting; for example, each class could be used to define a symbol indicating which groups do not differ from one another. The bottom section of Table 4 presents the final results for the Bonferroni, Holm, Hochberg, Hommel, and Shaffer procedures. In this example, Shaffer's procedure has greater power than both the Hochberg and Hommel procedures. This result is not unusual. In the cases examined thus far, Shaffer's procedure has uniformly been found to have the largest power. In many cases, the procedures only differ in the critical p-value for testing, and yield the same set of significant differences. In the cases in which the procedures have yielded different results, Shaffer's procedure has uniformly rejected a larger number of hypotheses than both the Hochberg and Hommel procedures.

5.2. Example 2

Example 2 represents a much more difficult problem, all pairwise comparisons of the jurisdictions participating in the 1994 NAEP TSA of reading. A representative sample of fourth-grade students attending pubic schools was selected in each jurisdiction. Students were then administered sections of the NAEP reading assessment, and the results were combined to estimate the average reading proficiency of students within the state, and for various demographic subgroups within the state. See Williams, Reese, Campbell, Mazzeo, and Phillips (1995) for more detailed presentation of the results, and Mazzeo, Allen, and Kline (1995) describes the sampling and psychometric procedures used in deriving the results.

Forty-four jurisdictions participated and authorized the release of their results for public use. Table 5 presents the adjacency matrix for t-tests comparing the mean reading proficiency for each state. Rows and columns have been ordered by the state's mean reading proficiency. As in Table 4, "1" indicates comparisons which are significant using the unmodified Bonferroni critical value, and "X" indicates the additional comparisons deemed significant using Shaffer's procedure.

Table 6 lists the undifferentiated classes based on the Bonferroni critical value. This example gives some sense of the complexity of implementing the S2 procedure. There are 20 classes, and they overlap one another substantially. The bottom section of Table 6 compares several procedures. As in Example 1, Shaffer's procedure yields more power, rejecting 17 more hypotheses than Hochberg's procedure and 9 more than Hommel's.

Table 5: Adjacency matrix for Example 2: Comparison of reading proficiency of 44 jurisdictions

	ME	ND	WI	NH	MA	IA	CT	MT	WY	NE	RI	IN	NJ	MN	DD	UT	MO	PA	NC	CO	VA	WV	ID	WA	TN	TX	MI	NY	KY	MD	AR	AL	GA	DE	AZ	FL	NM	SC	MS	HI	CA	LA	GU	DC
ME (228)	—	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ND (225)	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WI (224)	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NH (223)	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MA (223)	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IA (223)	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CT (222)	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MT (222)	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WY (221)	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NE (220)	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RI (220)	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IN (220)	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NJ (219)	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MN (218)	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DD (218)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UT (217)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MO (217)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PA (215)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NC (214)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CO (213)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VA (213)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WV (213)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ID (213)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
WA (213)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TN (213)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TX (212)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MI (212)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NY (212)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KY (212)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MD (210)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AR (209)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0	0
AL (208)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0	0
GA (207)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0	0
DE (206)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0	0
AZ (206)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0	0
FL (205)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0	0
NM (205)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0	0
SC (203)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0	0
MS (202)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0	0
HI (201)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0	0
CA (197)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0	0
LA (197)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0	0
GU (181)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—	0
DC (179)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	—

Table 6: Undifferentiated Classes and Family Sizes Example 2: Comparison of Reading Proficiency of 44 Jurisdictions

Undifferentiated Classes (Bonferroni)

$N_1 = 9$, E_1 = {ME, ND, WI, IA, MA, NH, MT, CT, WY}
$N_2 = 13$, E_2 = {ND, WI, IA, MA, NH, MT, CT, WY, RI, IN, NE, NJ, MN}
$N_3 = 14$, E_3 = {WI, IA, MA, NH, MT, CT, WY, RI, IN, NE, NJ, MN, UT, MO}
$N_4 = 15$, E_4 = {IA, MA, NH, MT, CT, WY, RI, IN, NE, NJ, MN, DD, UT, MO, PA}
$N_5 = 13$, E_5 = {MT, CT, WY, RI, IN, NE, NJ, MN, DD, UT, MO, PA, NC}
$N_6 = 14$, E_6 = {CT, WY, RI, IN, NE, NJ, MN, DD, UT, MO, PA, NC, MI, TX}
$N_7 = 14$, E_7 = {WY, RI, IN, NE, NJ, MN, DD, UT, MO, PA, NC, VA, MI, TX}
$N_8 = 18$, E_8 = {RI, IN, NE, NJ, MN, DD, UT, MO, PA, NC, CO, WV, VA, MI, TX, ID, WA, TN}
$N_9 = 18$, E_9 = {IN, NE, NJ, MN, DD, UT, MO, PA, NC, CO, WV, VA, MI, TX, ID, WA, TN, KY}
$N_{10} = 16$, E_{10} = {MN, DD, UT, MO, PA, NC, CO, WV, VA, MI, TX, ID, WA, TN, KY, NY}
$N_{11} = 16$, E_{11} = {UT, MO, PA, NC, CO, WV, VA, MI, TX, ID, WA, TN, KY, NY, MD, GA}
$N_{12} = 16$, E_{12} = {MO, PA, NC, CO, WV, VA, MI, TX, ID, WA, TN, KY, NY, MD, AR, GA}
$N_{13} = 17$, E_{13} = {PA, NC, CO, WV, VA, MI, TX, ID, WA, TN, KY, NY, MD, AR, AL, GA, AZ}
$N_{14} = 16$, E_{14} = {CO, WV, VA, MI, TX, ID, WA, TN, KY, NY, MD, AR, AL, GA, AZ, FL}
$N_{15} = 16$, E_{15} = {VA, MI, TX, ID, WA, TN, KY, NY, MD, AR, AL, GA, AZ, DE, FL, NM}
$N_{16} = 12$, E_{16} = {MI, TX, KY, MD, AR, AL, GA, AZ, DE, FL, NM, SC}
$N_{17} = 11$, E_{17} = {MD, AR, AL, GA, AZ, DE, FL, NM, SC, MS, HI}
$N_{18} = 8$, E_{18} = {GA, AZ, FL, NM, SC, MS, HI, CA}
$N_{19} = 8$, E_{19} = {GA, FL, NM, SC, MS, HI, CA, LA}
$N_{20} = 2$, E_{20} = {GU, DC}

Bonferroni: $t_\ell = 252$

Final: $t_\ell = 237$

Procedure	P_{critical}	Number of hypotheses not rejected
Bonferroni	.000053	457
Holm	.000113	442
Hochberg	.000113	442
Hommel	.000141	434
Shaffer	.000211	425

6. Conclusion

This paper has presented some theoretical results, and an algorithm, which allow Shaffer's S2 procedure for multiple comparisons to be applied to a larger number of groups than has been possible in the past. Westfall's (1997) method for implementing the S2 procedure, although more general than the one presented here, is feasible for relatively a small number of groups; $G = 6$ or 7 is probably the maximum. A commercially available software implementation of Shaffer's procedure, MultComp (Prosoft, 1994), restricts the number of groups to 10, and appears to implement only Shaffer's S1 procedure. The present approach implements the more powerful S2 procedure. Earlier work by Rasmussen (1993) assumed strict transitivity of statistical tests (as would be the case in comparing several means with equality of variances). The present method requires no such assumption.

This paper demonstrates an application of the procedure with 44 groups. In addition to the 1994 NAEP Reading Trial State Assessment, the results of four other NAEP TSA data sets have been successfully analyzed, the largest consisting of 45 jurisdictions. The computational burden was minimal; the longest run required approximately 15 seconds to run on a 133 MHz Pentium-based machine. The current version of the program is dimensioned to handle 100 groups, although the NAEP TSA comparisons are the largest problems that have been run to date.

Shaffer's procedure was developed in terms of pairwise comparisons, but, with some modification, it may be fruitfully adapted to some other situations. Unlike other recent approaches (e.g., the FDR approach of Benjamini and Hochberg, 1995), FWER control is maintained. Shaffer's procedure has the advantage of making no assumptions beyond those in the Bonferroni procedure, and is, therefore completely general. In the case of pairwise comparisons, it can be substantially more powerful than both the original Bonferroni procedure and other sequential procedures such as that of Holm. Results presented in Examples 1 and 2 indicate that, at least in some circumstances, Shaffer's procedure is also more powerful than both the Hochberg and Hommel procedures.

The only cost of applying Shaffer's S2 procedure is that of (previously overwhelming) complexity of the analysis. The present paper has provided methods whereby the complexity may be easily handled, making the procedure usable for problems that were previously out-of-reach. The success in applying the method to the NAEP Trial State Assessment data bears witness to the usefulness of the results presented here.

Acknowledgments

The work reported herein was supported, in part, under the National Assessment of Educational Progress (Cooperative Agreement No. R999G50001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education, and in part by ETS Research Allocation Project 794-10.

Portions of the research presented in this paper were presented at the annual meeting of the Psychometric Society, Gatlinburg, TN, June 1997.

The author would like to thank Juliet Shaffer, James Carlson, Erich Lehmann, the editor, and two anonymous reviewerws for numerous helpful comments. Thanks also go to David Freund of ETS for preparation of the datasets from the National Assessment of Educational Progress Trial State Assessments.

References

[1] Benjamini, Y, & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57** 289–300. MR1325392

[2] Bretz, F., Hothorn, T. and Westfall, P. (2002). On multiple comparisons in R. *R News*, **2**(3), 14–17. [Available online at http://cran.r-project.org/doc/Rnews]

[3] Cliff, N. (1975). Complete orders from incomplete data: Interactive ordering and tailored testing. *Psychological Bulletin*, **82** 289–302.

[4] Cliff, N. (1992). PAIRDEL.BAS: Program for computing matched-data d-statistics [computer program]. Los Angeles, CA: Psychology Department, University of Southern California.

[5] Cliff, N. (1993). Dominance relations: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, **114** 494–509.

[6] Donoghue, J. R. (1995, April). *Within-group covariance and variable weighting in cluster analysis: Extension to the P-variable case*. Paper presented at the annual meeting of the American Educational Research Association in San Francisco, CA.

[7] Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *J. Amer. Statist. Assoc*, **87** 162–170. MR1158635

[8] Finner, H. and Roters, M. (2002). Multiple hypothesis testing and expected number of Type I errors. *Ann. Statist.*, **30** 220–238. MR1892662

[9] Grechanovsky, E. and Hochberg, Y. (1999). Closed procedures are better and often admit a shortcut. *J. Statist. Plann. Inference*, **76** 79–91. MR1673341

[10] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75** 800–802. MR995126

[11] Hochberg, Y. and Rom, D. M. (1995). Extensions of Simes' test for logically related hypotheses. *J. Statist. Plann. Inference*, **48** 141–152. MR1366786

[12] Hochberg, Y. and Tamhane, A. (1987). *Multiple comparison procedures*. New York: Wiley. MR914493

[13] Holland, B. S. and Copenhaver, M. D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics*, **43** 417–423. MR897410

[14] Holland, B. S. and Copenhaver, M. D. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin*, **104** 145–149.

[15] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, **6** 65–70. MR538597

[16] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, **75** 383–386.

[17] Hsu, J. C. (1996). *Multiple comparisons: Theory and Methods*. New York: Chapman and Hall. MR1629127

[18] Hubert, L. and Arabie, P. (1985). Comparing partitions. *J. Classification*, **2** 193–218.

[19] Liu, W. (1996). Multiple tests of a non-hierarchical family of hypotheses. *J. Roy. Statist. Soc. Ser. B*, **58** 455–461. MR1377844

[20] Mazzeo, J., Allen, N. L. and Kline, D. L. (1995). *Technical report of the NAEP 1994 Trial State Assessment program in reading*. Washington, DC: Office of Educational Research and Improvement, U. S. Department of Education.

[21] Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63** 655–660. MR468056

[22] Prosoft (1994). *MultComp 2.1 for PC* [computer program]. Philadelphia, PA: Author.

[23] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc*, **66** 846–850.

[24] Rasmussen, J. L. (1991). SHAFHC: A FORTRAN implementation of Shaffer's multiple comparison procedure with HC enhancement. *Psychometrika*, **56** 153–154.

[25] Rasmussen, J. L. (1993). Algorithm for Shaffer's multiple comparison tests. *Educational and Psychological Measurement*, **56** 329–335.

[26] Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, **77** 663–665. MR1087860

[27] Sarkar, S. K. (1998). Some probability inequalities for ordered MTP2 random variables: A proof of the Simes conjecture. *Ann. Statist.*, **26** 494–504. MR1626047

[28] Sarkar, S. K. and Chang, C. K. (1997). Simes' method for multiple hypothesis testing with positively dependent test statistics. *J. Amer. Statist. Assoc*, **92** 1601–1608. MR1615269

[29] Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc*, **81** 826–831.

[30] Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, **46** 561–584.

[31] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73** 751–754. MR897872

[32] Toothaker, L. E. (1991). *Multiple comparisons for researchers*. Newbury Park, CA: Sage.

[33] Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, **5** 99–114. MR30734

[34] Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *J. Amer. Statist. Assoc.*, **92** 299–306. MR1436118

[35] Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D. and Hochberg, Y. (1999). *Multiple comparisons and multiple tests using the SAS® system*. Cary, NC: SAS® Institute Inc. Books by Users.

[36] Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: Wiley.

[37] Williams, P. L., Reese, C. M., Campbell, J. R., Mazzeo, J. and Phillips, G. W. (1995). *NAEP 1994 Reading: A first look (Revised Edition: October, 1995)*. Washington, DC: Office of Educational Research and Improvement, U. S. Department of Education.