NISS

Controlling Error in Multiple Comparisons, with Special Attention to the National Assessment of Educational Progress

Valerie S. L. Williams, Lyle V. Jones, and
John W. Tukey

Technical Report Number 33
December, 1994

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Controlling Error in Multiple Comparisons, with Special Attention to the National Assessment of Educational Progress*

Valerie S. L. Williams
National Institute of Statistical Sciences

Lyle V. Jones
The University of North Carolina at Chapel Hill

John W. Tukey
Princeton University

ABSTRACT

Adjustment procedures for multiplicity are investigated, including the traditional Bonferroni technique, a sequential Bonferroni technique developed by Hochberg (1988), and a sequential approach proposed by Benjamini and Hochberg (in press) for controlling the false discovery rate (FDR). The procedures are illustrated and compared, based on examples from several data sets, including the National Assessment of Educational Progress (*NAEP*) and the *NAEP* Trial State Assessment. One advantage of the FDR procedure is shown to be its consistency about the statistical significance of comparisons over alternative choices of family size. Simulation studies show that all three procedures maintain a false discovery rate bounded above by α (or $\alpha/2$). For both uncorrelated and pairwise families of comparisons, the FDR technique results in greater power than that for the Hochberg or Bonferroni procedures and the power advantage of the FDR procedure increases with the number of comparisons. We recommend that in reporting results from the Trial State Assessment, *NAEP* discontinue use of the Bonferroni procedure in favor of the FDR technique.

---

* Research supported through the National Science Foundation Grant No. DMS-9208758 and NCES/NSF through NSF RED-9350005. The authors are grateful to Susan Ahmed, Robert Burton, and others at the National Center for Education Statistics for their help in framing the issues, and to Jerome Sacks, Juliet P. Shaffer, and David Thissen for constructive suggestions. Special thanks are directed to Christopher Wiesen who provided critical support in developing software for the graphic displays and simulation studies.

Controlling Error in Multiple Comparisons,
with Special Attention to the National Assessment of Educational Progress

"If enough statistics are computed, some of them will be sure to show structure."
— Diaconis (1985)

Two questions often asked about one, or several, observed comparisons are: (a) whether we should be confident about the direction — the sign — of the corresponding underlying population comparison, and (b) for what interval of values should we be confident that it contains the value for the population comparison. Most often, each comparison will be a simple difference between two separately estimated quantities. The present report focuses on (a), above, in the interest of presenting some basic issues as starkly as possible. It expands on concepts introduced by Tukey (1991, 1993) and by Benjamini and Hochberg (in press).

Assume that we need statistical procedures to control some kind of error rate at a conventional value ($\alpha = .05$, perhaps). For a single comparison, $\alpha/2$ provides a bound related to the probabilities of combinations of

- deciding with confidence that a population comparison goes in one direction
- while the corresponding population comparison actually goes in the opposite direction (has the opposite sign).

This formulation assumes, as experience has taught us, that no population comparison is exactly zero (to many decimal places).

For a given critical value, the probability α, considered as a function of the (never zero) value of the population comparison, approaches a flat maximum when the population comparison approaches zero. Therefore, the conventional emphasis on "the null hypothesis" is not surprising. The importance of the null hypothesis is not that it is *null*, but rather that

- as a limit, it is the least favorable case, and
- situations with small non-zero values of the population comparison ("perinull" situations) behave much as if they were at that limit.

The probability of erroneous confidence, defined more generally than above, is discontinuous by a factor of two depending on whether the population comparison is zero or near-zero. This is so because confidence in either of the two directions is erroneous if a population comparison is exactly zero, but only one direction is erroneous when the population comparison is not precisely zero; also, so long as the true difference is close to zero, values beyond the selected critical value are very nearly as likely to be of one sign as the other.

The probability, α, is the probability of the traditional "Type I error." Frequently — and rather misleadingly — it is considered to be "the probability of deciding to be confident about the direction of an observed comparison when the population difference is exactly zero." Instead, we recommend thinking of α, in the simplest case, as "a bound on twice the

probability of being erroneously confident about the direction of the population comparison."

Multiplicity arises in situations where more than one comparison is assessed. Unless some correction is incorporated, the overall (simultaneous) Type I error rate — the probability that the decision for any one or more comparison will be in error — will exceed (often very substantially) the nominal α (which still would apply to any single comparison whenever that comparison can be taken alone). When a number of comparisons are assessed together, it is appropriate — and usually essential — to adjust for the increased probability of (simultaneous) Type I error, i.e., the probability of finding at least one erroneous confident direction. See Shaffer (1994) for a current review of the range of multiple comparison adjustments that have been proposed to control one kind or another of an overall Type I error rate.

The Bonferroni adjustment is a simple and trustworthy statistical procedure for assuring simultaneously that the probability of any Type I error is no greater than α. However, in the trade-off between the control of Type I error and statistical power, power is severely reduced when the *simultaneous* error rate is made no greater than α by the use of the Bonferroni adjustment.[1]

Two sequentially-rejective techniques described in Benjamini and Hochberg (in press) provide greater statistical power than the Bonferroni correction while still attempting to control the rate of erroneous declarations of confidence. One of these, the Hochberg procedure (Hochberg, 1988), controls the family-wise error rate at α, which is then a bound on the probability of making any (one or more) Type I error in a given family of comparisons; this bound is very nearly sharp when all population comparisons are zero. In contrast, the Benjamini and Hochberg False Discovery Rate (FDR) technique attempts to control the fraction of false discoveries, roughly, the average fraction, among all confident directions asserted, that are errors; therefore, $\alpha/2$ provides an approximate bound for a given family of comparisons on the expected value of the ratio of (a) the *number of erroneous declarations* of confident differences to (b) the *total number of declarations* of confidence *plus 1*. (See Appendix A for discussion of false discovery rates and false discovery proportions, especially near the null hypothesis.)

Let p_{crit} be the tail area (usually for each of two tails) of the null sampling distribution of the test statistic for any single comparison being judged by a multiplicity-respecting procedure. Each procedure will stipulate a probability of error or average fraction of error that is bounded by α when assessing confidence (or nonconfidence) of direction. The value of p_{crit} will depend on the sort of confidence to be attained. Let m be the number of comparisons and $i = 1, ..., m$ be the rank of the p-value of the statistic for the comparison concerned when ordered from smallest to largest, so that the observed p-values — $p_{(i)}$ for the i^{th} comparison — are weakly increasing from $i = 1$ to $i = m$. Four distinct approaches are defined as follows:

Bonferroni: the traditional Bonferroni-adjusted (two-tailed) p-value; the critical value of the statistic is such that $p_{crit} = p_{BON} = \alpha/2m$ in each tail of the distribution of that statistic.

Hochberg: according to Hochberg (1988), be confident of the observed direction of the

i^{th} comparison when, beginning with $i = m$ and continuing toward $i = 1$, $p_{(i)} \leq p_{crit} = p_{HOC}(i) = \alpha/2(m-i+1)$; then stop and declare a confident direction for all comparisons for which $j \leq i$. Thus, $p_{HOC}(i) = m/(m-i+1)p_{BON}$.

FDR: according to Benjamini and Hochberg (in press), be confident of the observed direction of the i^{th} comparison when, beginning with the m^{th} comparison, $p_{(i)} \leq p_{crit} = p_{FDR}(i) = i\alpha/2m$; then stop and declare a confident direction for all comparisons for which $j \leq i$; thus, $p_{FDR}(i) = ip_{BON}$.

Unadjusted: the unadjusted test; be confident if $p_{(i)} \leq p_{crit} = p_{UNA} = \alpha/2$. Thus, $p_{UNA} = mp_{BON}$.

Because

$$1 \leq m/(m-i+1) \leq i \leq m ,$$

it must be true that

$$p_{BON} \leq p_{HOC} \leq p_{FDR} \leq p_{UNA} ;$$

the four collections of confident directions corresponding to these four approaches are nested, with the unadjusted p_{crit} value the largest and the Bonferroni smallest.

Each of the four procedures operates in terms of a critical area, p_{crit}. For any data set involving standard errors based upon a single number of degrees of freedom, p_{crit} can be directly translated, first into a single critical t-value, t_{crit}, and then, if all standard errors are the same, into a single critical effect size, d_{crit}, such that

$$|y_i - y_j| \geq d_{crit}$$

is the condition for confidence in direction.

To illustrate the calculations called for by the four procedures, we present those calculations in *An Election Example*, for the Main Effects for Election Year (below). Because in this example a common standard error is used, there is a unique critical effect size, d_{crit}, which corresponds to a vertical line in Figure 1 and Figure 2. In this simple situation, the four vertical lines, one for each procedure (shown in the figures by the four edges of two gray stripes) tell almost the whole story.

An Election Example

This illustration applies the four techniques to an example involving multiple comparisons, taken from Tukey, Mosteller, and Hoaglin (1991). The data are the proportions of state-by-state vote for Franklin D. Roosevelt in the four U.S. Presidential elections from 1932 to 1944 for 39 states. The states were organized by Tukey, Mosteller, and Hoaglin into 13 Groupings of three geographically contiguous states each, and they assessed the main effects for Election Year and for Grouping, and the effects for State nested within Grouping. Here, we perform similar analyses, and we replicate, as well, an analysis of Election Year by Grouping interaction performed by Tukey and Hoaglin (1991). In all cases, we set $\alpha = .05$.

Main Effects for Election Year

The first set of pairwise tests involves six comparisons of four presidential election years — 1932, 1936, 1940, and 1944 — each with each other. For these elections, the percentage of vote for Roosevelt was 63%, 65%, 59%, and 56%, respectively.

To compare rates of confident decisions about the direction of differences, t-statistics were computed using the standard errors of the differences between Election Years (based on the Election Year × Grouping interaction) and the degrees of freedom provided by Tukey and Hoaglin (1991, p. 339). The comparisons are ordered from $i = 6$ (largest p-value, smallest absolute value of t) to $i = 1$ (smallest p-value, largest absolute value of t), as presented in Table 1.

Table 1.
Critical values of p_{UNA}, $p_{FDR}(i)$, $p_{HOC}(i)$, and p_{BON} for the Main Effects for Election Year from the *Election Example* ($df = 36$).

Comparison (i)	t	p-value	p_{UNA}	$p_{FDR}(i)$	$p_{HOC}(i)$	p_{BON}
1932 – 1936 (*6*)	−1.49	.07152	.025	.02500	.02500	.00417
1940 – 1944 (*5*)	1.63	.05556	.025	.02083	.01250	.00417
1932 – 1940 (*4*)	2.79	.00417	.025*	.01667*	.00833*	.00417*
1936 – 1940 (*3*)	4.29	.00006	.025*	.01250*	.00625*	.00417*
1932 – 1944 (*2*)	4.42	.00004	.025*	.00833*	.00500*	.00417*
1936 – 1944 (*1*)	5.92	.00000	.025*	.00417*	.00417*	.00417*

The critical values for each approach — unadjusted, FDR, Hochberg, and Bonferroni — are shown in the respective columns, p_{UNA}, $p_{FDR}(i)$, $p_{HOC}(i)$, and p_{BON}. Recall that $p_{UNA} = \alpha/2$, and $p_{BON} = \alpha/2m$, regardless of i. The FDR and the Hochberg techniques are both "step-up" procedures; for each, we start by testing the least significant comparison (here, $i = 6$) and then work toward the most significant.

With the Hochberg technique, beginning at $i = m = 6$ comparing 1932 and 1936, the p-value(*6*) = .07152 > $p_{HOC}(6) = \alpha/2(m-i+1) = .025$, so we may not infer a confident direction of difference between 1932 and 1936. For the comparison of 1940 and 1944, $i = m-1 = 5$, the observed p-value, again, is greater than the critical p-value for the Hochberg technique, p-value(*5*) = .05556 > $p_{HOC}(5) = .0125$, and we may not be confident at, $\alpha = .05$, of the direction of the difference between 1940 and 1944. For the 1932–1940 comparison, $i = m-2 = 4$, the p-value(*4*) = .00417 < $p_{HOC}(4) = .00833$; we conclude with confidence that the proportion of the vote for Roosevelt in 1932 was sufficiently greater than the proportion in 1940 to be attributable to other than chance variability. At this point, we end the Hochberg procedure without testing any further comparisons; all remaining comparisons in the specified family,

together with the one last tested, comprise confident directions. Because, for percentage of vote for Roosevelt, $|difference| \geq 4\%$ leads to confidence about the direction of the difference of the vote while $|difference| \leq 3\%$ leads to lack of confidence, nonconfidence of direction can be separated from confidence of direction anywhere between 3% and 4% of the vote, with 3.5% a natural choice.

Applying the FDR technique, we again begin with $i = m = 6$ for the 1932–1936 comparison. The observed p-value for this comparison, $p\text{-value}(6) = .07512 > p_{FDR}(6) = i\alpha/2m = .025$; therefore, we are not confident about the direction of the difference between 1932 and 1936. Next, for the comparison between 1940 and 1944, $i = m-1 = 5$, we are not confident about the direction of the difference between 1940 and 1944 because the $p\text{-value}(5) = .05556 > p_{FDR}(5) = .02083$. We test next the comparison between 1932 and 1940, $i = m-2 = 4$; because the $p\text{-value}(4) = .00417 < p_{FDR}(4) = .01667$, we are confident of the direction of this difference. We conclude, as for the Hochberg technique, that this difference ($i = 4$), together with the remaining comparisons ($i = 3, 2, 1$), represent confident directions of differences.

For this example, all four test procedures yield the same conclusions, which also agree with those reported in Tukey, Mosteller, and Hoaglin (1991) using the Bonferroni: We are confident that Roosevelt's percentage of the vote in both 1932 and 1936 was sufficiently higher than that in 1940 and 1944 to be attributable to other than chance variation. However, the vote for Roosevelt in 1932 did not differ in a confident direction from that of 1936, nor did the 1940 vote differ in a confident direction from that of 1944. (All this assumes that the error term reflects the sort of variability that should be included in the standard errors.)

Main Effects for Grouping

A second set of tests compares each of the 13 Groupings of states with each other Grouping. Here, t-statistics for each of 78 (= 13×12/2) paired comparisons are calculated. The numbers of confident decisions about direction are found to be 29 for the unadjusted approach, 20 for the FDR procedure, 12 for the Hochberg technique, and 11 using the Bonferroni procedure. Each of the discrepancies between the Bonferroni and the FDR adjustments involve comparisons of one of the Groupings of southern states — Grouping #8 (GA, NC, SC), #9 (AL, KY, TN), or #10 (AR, MS, OK) — with one of the Groupings (from #1 to #7 or from #11 to #13) from other regions, reflecting the observed pattern of stronger support for Roosevelt in the South. Figure 1 represents these comparisons graphically:

- on the vertical scale is the average percent of the vote for Roosevelt for each Grouping;
- the two 45° lines emanating (to the right) from the location of each Grouping on the vertical scale are lightly drawn;
- there are 78 intersections of upward 45° and downward 45° lines, one representing each comparison;

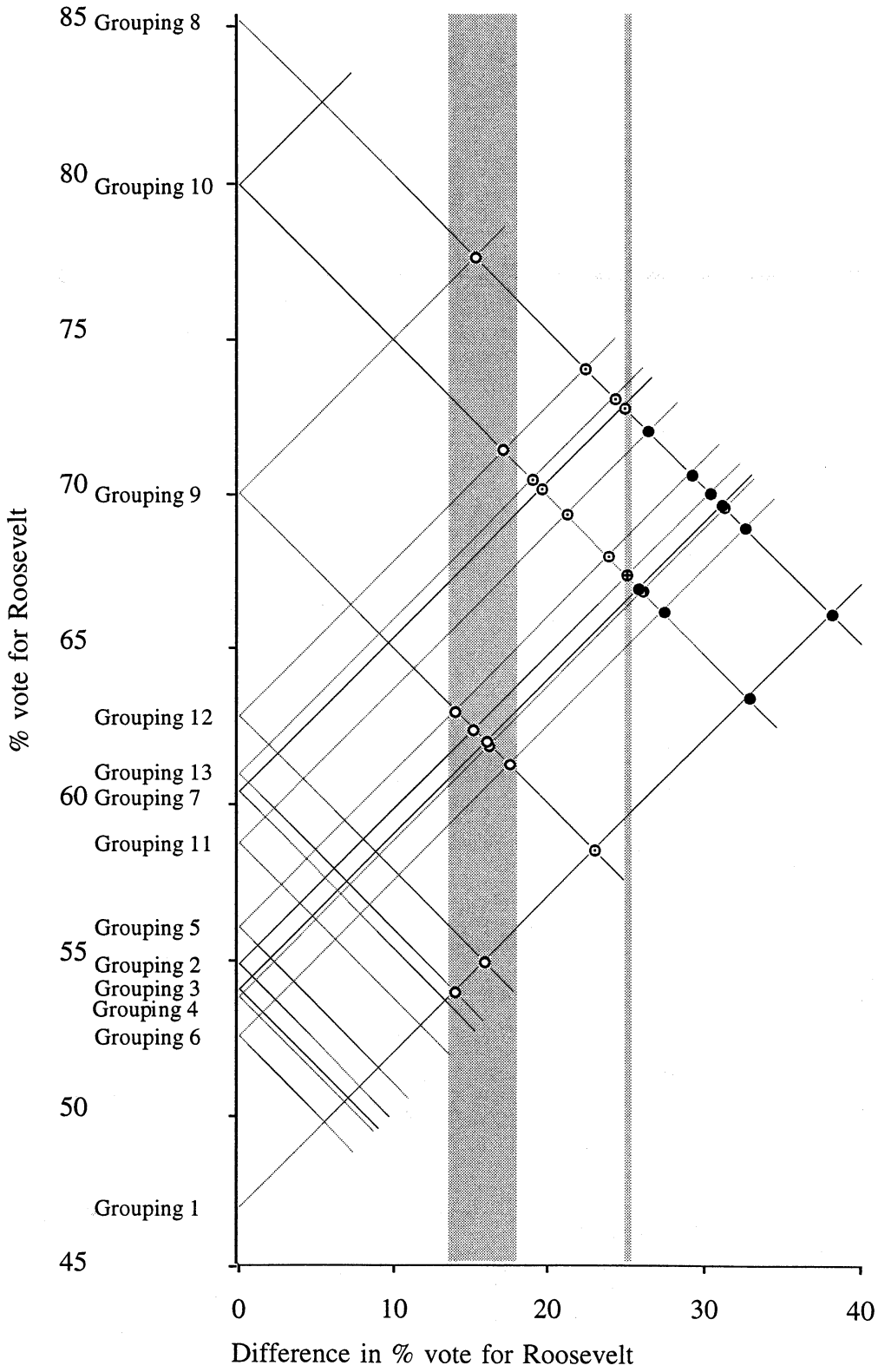

Figure 1.

Confident directions between Groupings: *Election Example* (*df* = 33.4).

Unadjusted = ○, FDR = ⊙, Hochberg = ⊕, Bonferroni = ●

Note: The broad gray stripe indicates Unadjusted, but not FDR, confidence; between the stripes: Unadjusted and FDR, but not Hochberg, confidence; the narrow gray stripe: Unadjusted, FDR, and Hochberg, but not Bonferroni, confidence.

- intersections are marked with solid circles (●) for comparisons that are Bonferroni-confident of direction (11 intersections);
- the lone intersection corresponding to Hochberg-confidence, but not Bonferroni-confidence, is marked with a circle and *plus* sign (⊕);
- the 8 intersections which are FDR-confident, but not Hochberg-confident are marked with a circle with a dot (⊙);
- the 9 additional intersections that are unadjusted-confident, but not FDR-confident, are marked with an open circle (○).

States within Grouping

Tukey and Hoaglin (1991; pp. 360-363) investigate differences between States within Grouping, taking into account the nested structure of the *Election Example* data (13 Groupings of 3 States each). Because their analysis utilizes the Studentized range for the triplets of states — so that they have only one statistic's value for each Grouping — they use $p_{BON} = .05/13$. The analysis presented here, however, involves mean differences between all within-Grouping pairs of states, three pairs within each Grouping, so $m = 39$ is required. Table 2 presents the findings in a format similar to that of Tukey and Hoaglin (1991, p. 362).

Table 2.
Confident directions between States within Groupings for the *Election Example*, by the Bonferroni procedure ($df = 78$). (Additional confident directions by the FDR are parenthesized.)

Grouping	Direction Bonferroni, (FDR)
1	New Hampshire >> Vermont (Maine)
5	Minnesota (North Dakota) >> South Dakota
6	Missouri >> Kansas (Nebraska)
7	Virginia >> Maryland, Delaware (Maryland >> Delaware)
8	South Carolina >> Georgia >> North Carolina
9	Alabama >> Tennessee >> Kentucky
10	Mississippi >> Arkansas >> Oklahoma
11	Montana >> Wyoming
12	Arizona >> New Mexico
Note:	No confident directions within Groupings 2, 3, 4, and 13.

For all 16 comparisons that are shown by unparenthesized state names, we are confident of the direction of differences using the Bonferroni adjustment; the FDR technique produces 4 additional confident directions for states shown in parentheses. In this case, the unadjusted per comparison approach was confident about 21 comparisons, and the FDR, 20;

the Hochberg technique was confident about 17 comparisons, and the Bonferroni adjustment, 16 (the same hypotheses rejected by the Studentized range analysis in Tukey and Hoaglin).

Tukey, Mosteller, and Hoaglin's (1991, pp. 18-19) analysis of pairwise differences between states, after subtraction of Grouping effects, was also replicated. This presentation involves 741 or 39×38/2 simultaneous comparisons for these residuals, displayed graphically in Figure 2.

The frequencies of confident directions for each of the four adjustment techniques are presented in Table 3. There is a striking increase in power associated with the FDR technique, but Hochberg's technique again, as anticipated, only slightly outperforms the Bonferroni adjustment. Of the confident directions by the unadjusted per comparison approach, more than 90% also are confident by the FDR procedure.

Table 3.
Number of comparisons of the residuals from the Grouping effect with confident direction, $m = 741$.

Unadjusted	FDR	Hochberg	Bonferroni
391	369	238	232

Election Year × Grouping Interaction Effects (E × G)

Interaction effects represent the failure of the effects of changes, one in each of the two factors, to be additive. What set of comparisons should we look at to describe and dissect interactions? A rather naive choice is to look at *all* double differences (cross differences), of which there are $rc(r-1)(c-1)/4$. We need pay less for multiplicity, however, by looking instead at pairwise comparisons (in one direction) of deviations from means (in the other direction). Equivalently, we can use interaction values defined in the usual way in place of the deviations.

The total number of conditional comparisons required is the number of "all pairwise comparisons within each column *and* ... all pairwise comparisons within each row" (Tukey & Hoaglin, 1991; p. 351). Two things must be considered carefully here: (i) "all pairwise comparisons within each column" refers to comparisons of *interactions* which reduce to $(y_{hj}-y_{ij})-(\bar{y}_h-\bar{y}_i)$, and (ii) because we are also looking at "all pairwise comparisons within each row," we are using both column and row comparisons to look at the same interactions. Here, in order to keep to an overall α of no more than 5%, we will spend 2.5% on each set of conditional comparisons of interactions, as Tukey and Hoaglin did. For these analyses, family size is defined as $m = 78 = 13(4\times3/2)$ for the pairwise comparisons within Grouping (rows), and $m = 312 = 4(13\times12/2)$ for the comparisons within Election Year (columns).

For the 78 pairwise comparisons within Grouping, the unadjusted per comparison

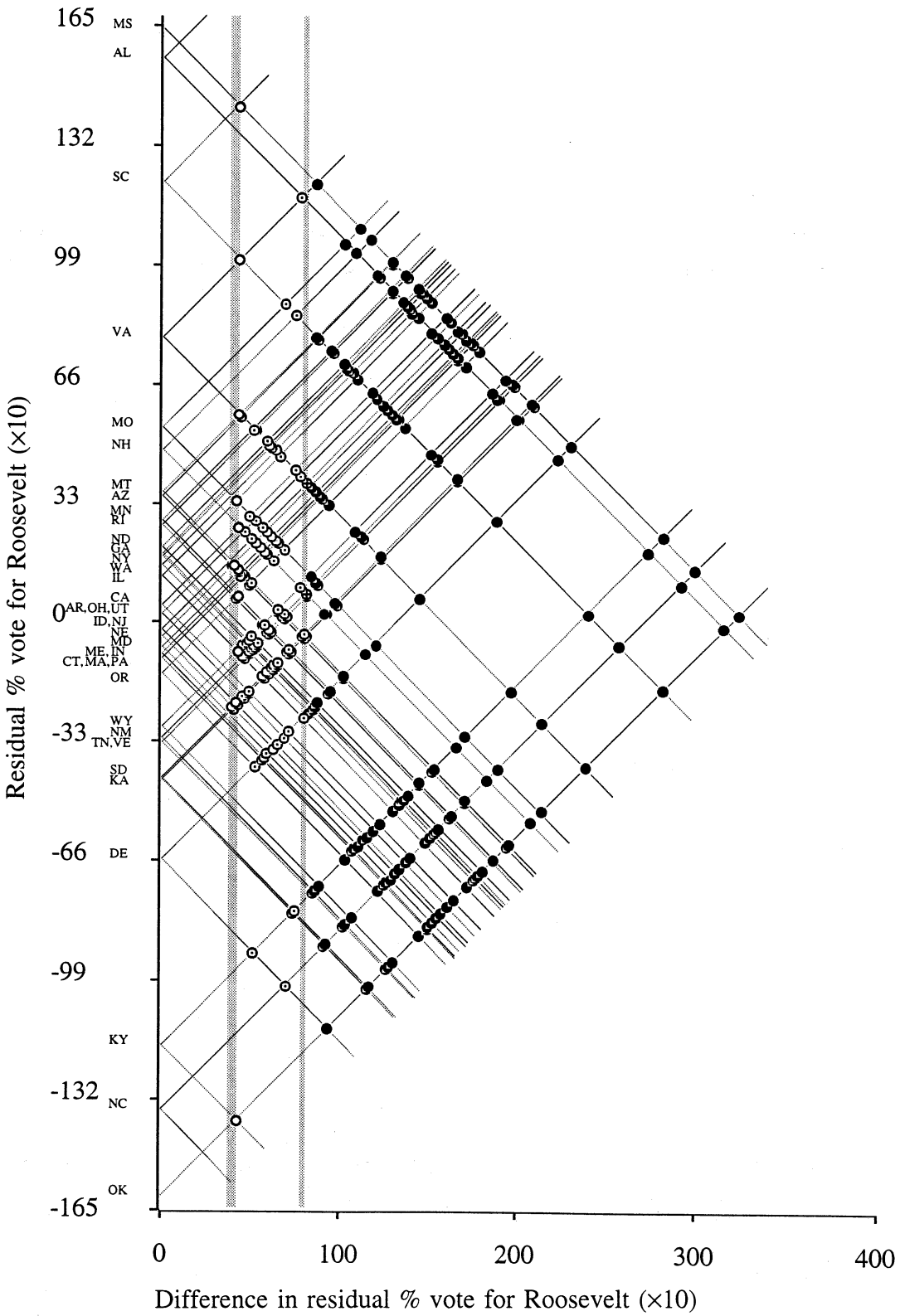

Figure 2.

Confident directions between States for residuals from the Grouping effect: *Election Example* (*df* = 78).

Unadjusted = ○, FDR = ⊙, Hochberg = ⊕, Bonferroni = ●

Note: The gray stripe on the left indicates Unadjusted, but not FDR, confidence; between the stripes: Unadjusted and FDR, but not Hochberg, confidence; the gray stripe on the right: Unadjusted, FDR, and Hochberg, but not Bonferroni, confidence.

approach results in a total of 20 confident directions, the FDR technique results in 12 differences with confident direction, and the Hochberg technique results in 10. With the more stringent p_{BON} = .00016 = .0125/78, the Bonferroni adjustment results in the detection of 9 significant differences within Grouping.

For the 312 comparisons within Election Year, the unadjusted per comparison approach produces 79 confident directions, while the FDR technique results in 52 confident directions. The Hochberg procedure detects 15 differences, and the Bonferroni adjustment produces the same 15 confident directions (p_{BON} = .00004 = .0125/312).

See Appendix B for a presentation of alternative portrayals of interaction effects for the *Election Example*.

A NAEP Trial State Assessment Example

Data from the National Assessment of Educational Progress (*NAEP*) Trial State Assessment (*TSA*) were subjected to an analysis parallel to the *Election Example* described above. Now, the appropriate standard error varies from one comparison to another. The data are average 8th-grade mathematics proficiency scores for the 34 states that participated in both the 1990 and 1992 *NAEP TSA* (Johnson, Mazzeo, & Kline, 1993; E. G. Johnson, personal communication, July 29, 1993).[2]

The states are classified into four geographical Regions of eight or nine states each: Central, Northeast, Southeast, and West. The four conditions of Type I error control are compared for the main effects for Year (1990 and 1992) and for Region, the effects for State nested within Region, and the Year × Region interaction effects, each with α = .05.

Main Effects for Year

Because there are only two Years of data — and therefore only one possible comparison of Years — it is not necessary to adjust for multiplicity. We are highly confident that average mathematics performance was better for 1992 8th-graders ($\bar{X}_{92}$ = 266.6) than for 1990 8th-graders ($\bar{X}_{90}$ = 263.4), t_{33} = 9.73. In fact, the 95% confidence interval for the mean improvement is from 2.6 to 3.8 scale score points, so that the average amount of increase is at least roughly known (almost known to one significant digit).

Main Effects for Region

To assess pairwise mean differences in mathematics performance for the four Regions, *t*-statistics were calculated for m = 6 comparisons (all pairwise differences) presented graphically in Figure 3. Regional mean scale scores are: Central $\bar{X}$ = 274.1, Northeast $\bar{X}$ = 266.9, West $\bar{X}$ = 263.7, and Southeast $\bar{X}$ = 256.4. The unadjusted per comparison approach and the FDR and Hochberg procedures each leads to confidence for five of the six differences — only the direction of the difference between the averages for the Northeast and the West Regions fails to reach confidence. The Bonferroni technique provides confidence for the

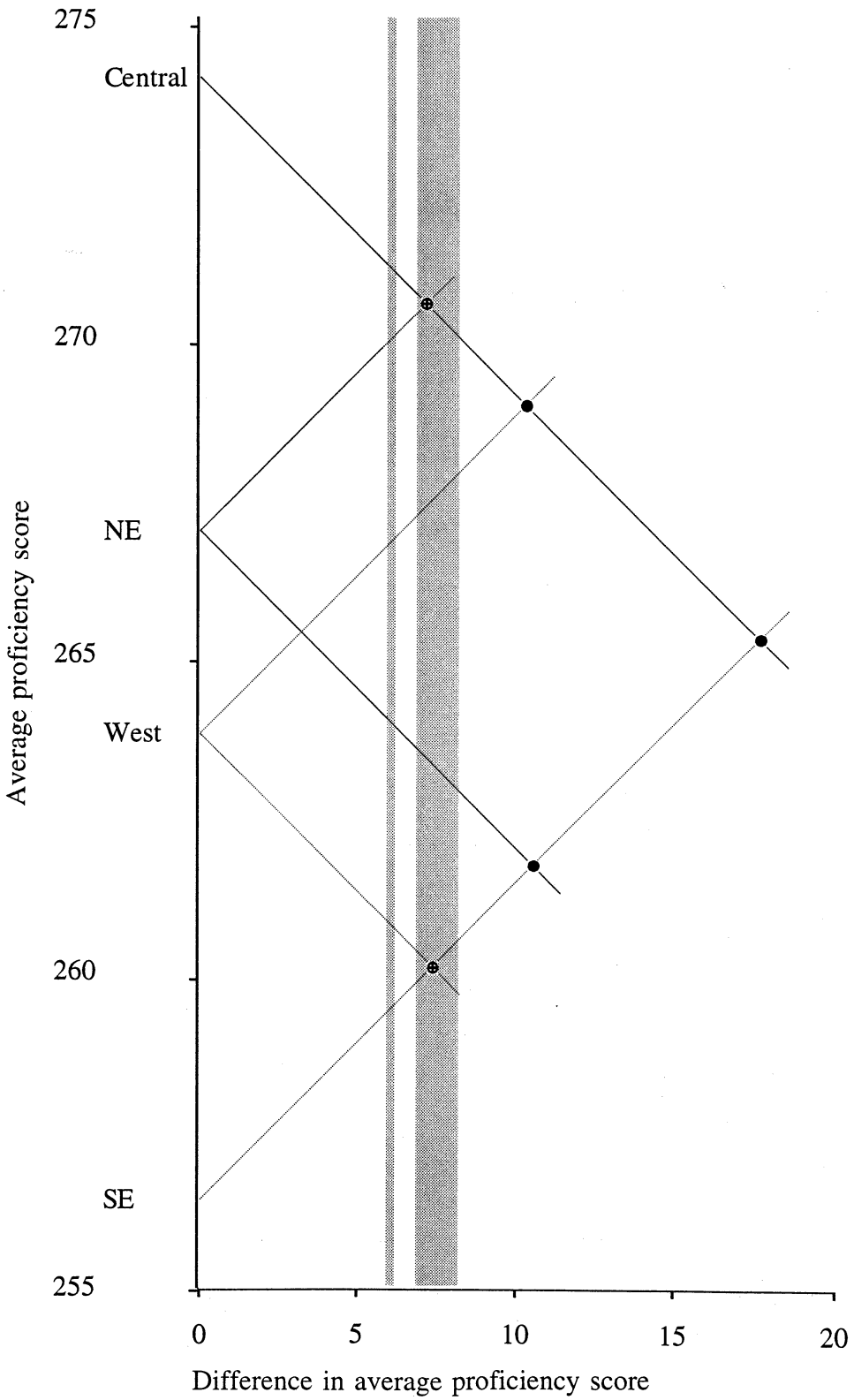

Figure 3.

Confident directions between Regions for the *NAEP TSA Example* (*df* = 30).

Unadjusted, FDR, and Hochberg = ⊕, Bonferroni = ●

Note: The narrow gray stripe indicates Unadjusted, but not FDR, confidence; between the stripes: Unadjusted and FDR, but not Hochberg, confidence; the broad gray stripe: Unadjusted, FDR, Hochberg, but not Bonferroni, confidence.

direction of only three of the six differences (Central *vs* Southeast, Central *vs* West, and Northeast *vs* Southeast).

States within Region

Differences among States within Region were investigated by considering all pairwise comparisons within Regions, $m = 128$ comparisons in all.[3] The results show considerable variation in student mathematics performance among States within Region: The unadjusted per comparison approach ($p_{UNA} = .025$) is confident about the direction of 100 differences, of which 96 are also found by the FDR technique, 74 are found by the Hochberg procedure, and 71 are found by the Bonferroni adjustment ($p_{BON} = .025/128 = .000195$). Table 4 shows the results from both the Bonferroni procedure (confident directions involve only states not in parentheses) and the FDR technique (confident directions are for states not in parentheses as well as for states enclosed in parentheses).

Table 4.
Confident directions between States within Region for the *NAEP TSA Example* by the Bonferroni procedure ($df = 30$, common standard error). (Additional confident directions by FDR are parenthesized.)

Region	Direction Bonferroni, (FDR)	Additional Directions by FDR
Central	ND >> (MN), NE, WI, IN, OH, MI	1
	IA >> (NE), (WI), IN, OH, MI	2
	MN, NE, WI >> IN, OH, MI	0
Northeast	NH >> (CT), (NJ), PA, NY, RI, MD, DE	2
	CT, NJ >> NY, RI, MD, DE	0
	PA >> (NY), RI, MD, DE	1
Southeast	VA >> KY, GA, FL, WV, AR, NC, AL, LA	0
	KY >> (AR), (NC), AL, LA	2
	GA >> (NC), AL, LA	1
	FL, WV >> (NC), (AL), LA	4
	AR >> (AL), LA	1
	NC, (AL) >> LA	1
West	WY, ID >> (CO), OK, AZ, TX, CA, NM, HI	2
	CO >> (OK), AZ, TX, CA, NM, HI	1
	OK >> (AZ), (TX), CA, NM, HI	2
	AZ >> (CA), (NM), HI	2
	TX >> (NM), HI	1
	(CA), (NM) >> (HI)	2

An alternative approach to assessing a different aspect of the behavior of States compared with the Region to which they belong involves computing deviations of States from their Regions, then comparing these residuals from a Region effect for all pairs of the 34

states. The frequencies of confident directions of difference among the 561 contrasts for each of the four multiple comparison techniques are presented in Table 5. Again, there is a very large increase in power associated with the FDR technique and the number of FDR rejections approaches the number found by the unadjusted per comparison approach. Figure 4 shows the 561 comparisons for the 34 states graphically.

Table 5.
Number of comparisons between States' residuals from the Region effect with confident direction, $m = 561$.

Unadjusted	FDR	Hochberg	Bonferroni
432	418	294	275

Year × Region Interaction Effects (Y × R)

The Year × Region interaction effects were not statistically significant when tested by an overall F-test.[4] Nevertheless, we applied the four multiple comparison procedures to illustrate their application to weakly-structured data. No cell mean differences were found to be statistically significant using any one of p_{UNA}, $p_{FDR}(i)$, $p_{HOC}(i)$, or p_{BON}.

Differences between All Pairs of States for 1992 8th-Grade Math Scores

All pairwise mean differences between the states' 1992 8th-grade mathematics achievement scores were compared. There were 41 states which participated in the 1992 assessment, resulting in a family size of $m = 41 \times 40/2 = 820$. Table 6 summarizes the number of confident directions and Figures 5a and 5b present the graphical comparison of the four multiplicity treatments. We use two pictures because, with different standard errors for different states, the boundaries are not straight lines — the use of two figures gives a clearer idea of what is happening. By the Bonferroni adjustment, there are 480 confident directions between states; the Hochberg technique admits 13 more confident directions, and the use of the FDR results in an additional 159. The unadjusted analysis increases the number of confident directions beyond the FDR technique by only 6.

Table 6.
Number of comparisons of differences between all pairs of States with confident direction, $m = 820$ (df taken as 60, individual state standard errors).[5]

Unadjusted	FDR	Hochberg	Bonferroni
658	652	493	480

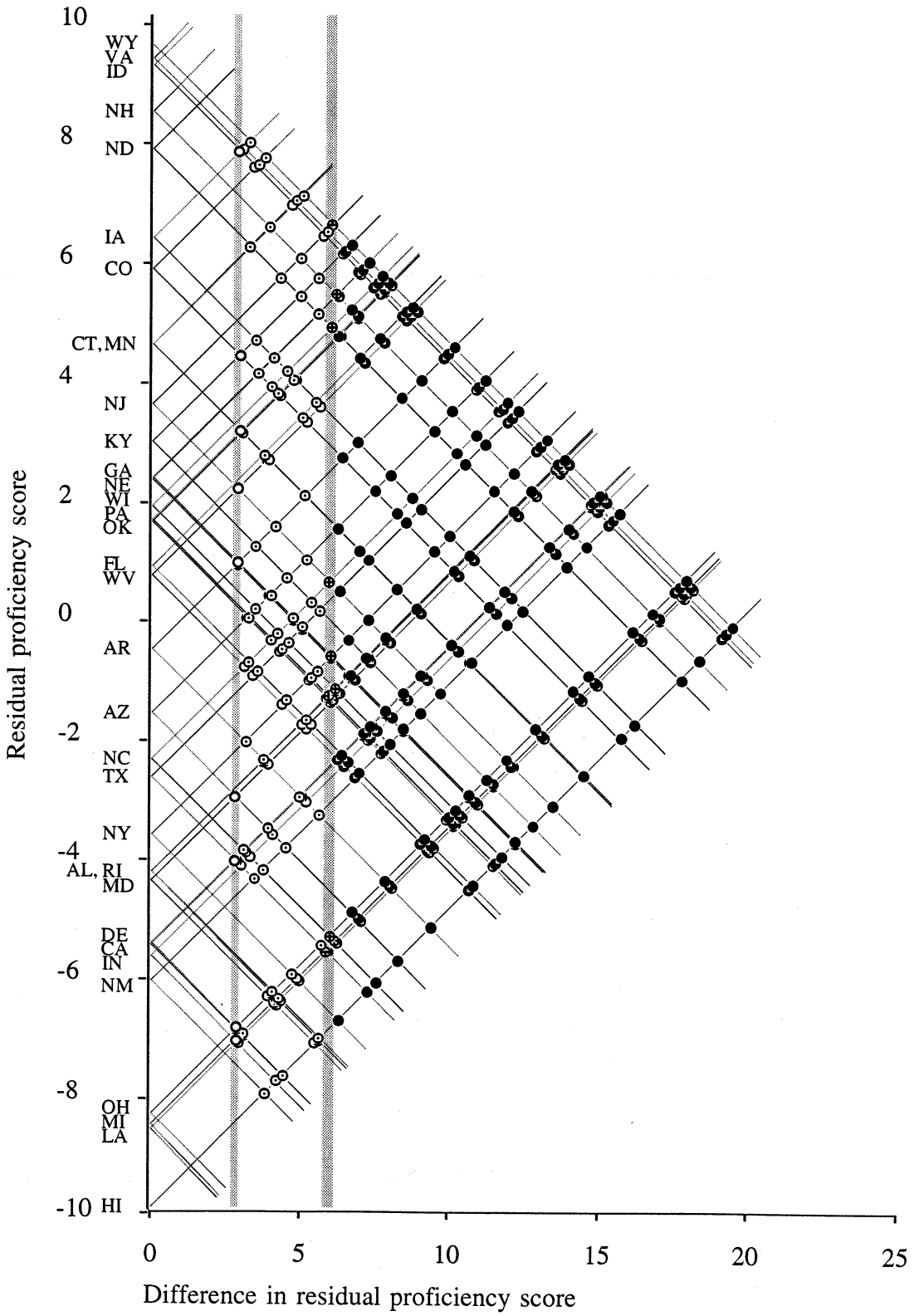

Figure 4.

Confident directions between States for residuals from the Region effect: *NAEP TSA Example* (*df* = 30).

Unadjusted = ○, FDR = ⊙, Hochberg = ⊕, Bonferroni = ●

Note: The gray stripe on the left indicates Unadjusted, but not FDR, confidence; between the stripes: Unadjusted and FDR, but not Hochberg, confidence; the gray stripe on the right: Hochberg, but not Bonferroni, confidence.

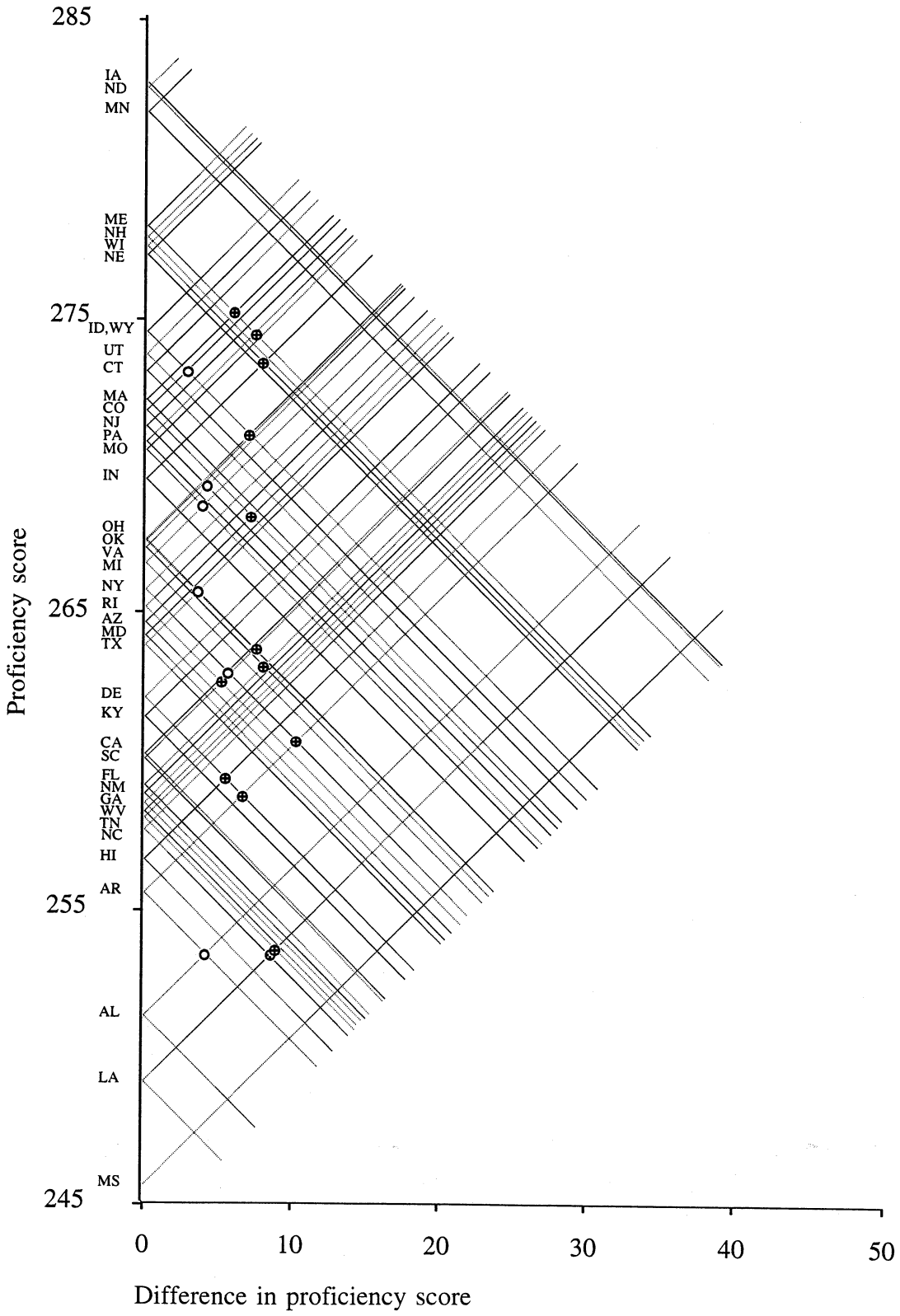

Figure 5a.

Confident directions between States, all pairwise comparisons: *NAEP TSA Example.*

Unadjusted = ○ and Hochberg = ⊕

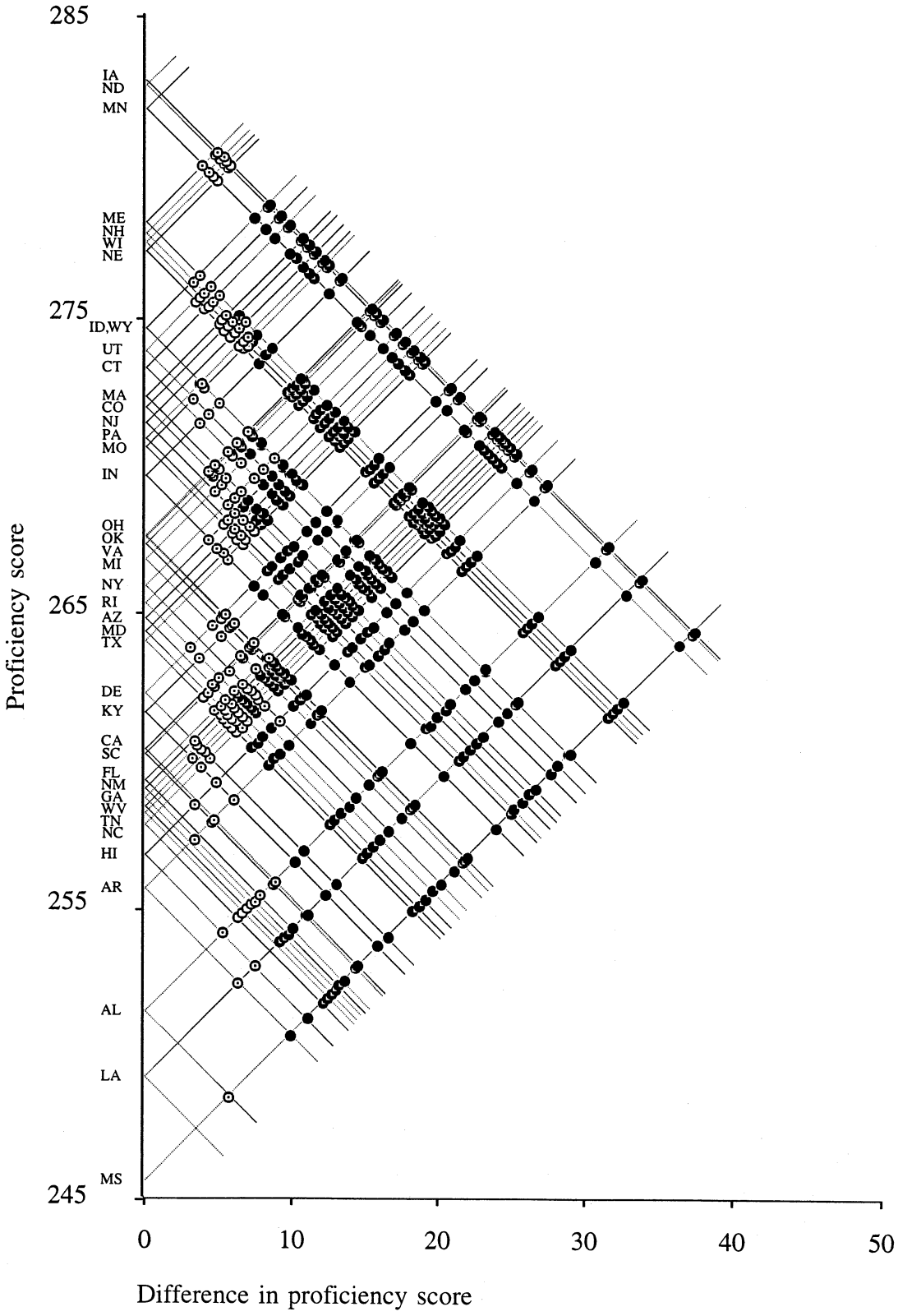

Figure 5b.

Confident directions between States, all pairwise comparisons: *NAEP TSA Example.*

FDR = ⊙ and Bonferroni = ●

Assessing State-by-State Change

One way of assessing year-to-year change in state mathematics achievement involves directly testing the significance of change in each of the 34 states. An ordinary Bonferroni adjustment for control of Type I error, as used by *NAEP*, results in an effective critical *p*-value of $p_{BON} = .000735$ for $\alpha = .05$. Table 7 contains the computed differences between 1990 and 1992 mathematics achievement means and the computed pooled standard errors. The table also includes the statistical decision about confidence in direction with confident direction indicated by an "*" in each column, for each state under the four multiplicity treatments. (For this and previous examples, we have chosen to present critical *p*-values as a basis for comparison of the adjustment procedures. The cutoff critical *t*-value for each procedure is also shown in the bottom line of Table 7. In Appendix C, we show comparisons based on critical values of t^2 rather than critical values of *p*.)

The unadjusted per comparison approach is confident of direction for 15 differences, whereas the FDR procedure is confident for 11 of these, and Hochberg's technique is confident of direction for 4 differences, the same 4 as the Bonferroni correction.

Assessing Differences between State Change and Average Change

The results for the main effects for Year show that the average mathematics proficiency improved substantially for the 34 states combined, $\bar{X}_{90} = 263.4$ and $\bar{X}_{92} = 266.6$. The state changes can also be evaluated with this average increase removed. In fact, when this change is subtracted out, there is no state for which we can be confident of differential change (from the average change) when an adjustment for multiplicity is made by using any one of the three procedures, p_{FDR}, p_{HOC}, or p_{BON}. (Using p_{UNA}, the drop in performance for Arkansas and Alabama is statistically significant, as are the two largest increases, those for Minnesota and North Carolina.) From a further analysis, we find that when using FDR with a more lenient $\alpha = .10$, there is still no state that confidently shows greater or lesser gain than the average of all 34 states. While we may confidently conclude an average gain for the participating states, we cannot be confident that any state gained more or less than the average.

A NAEP Example Bearing on the Consistency of Findings over Differing Family Sizes

The determination of family size, *m*, is critical to multiple hypothesis testing. Family size is always *the number of contrasts under consideration.* However, there may be legitimate ambiguities about family size for a particular set of data. A desirable feature for a multiple comparison procedure is that it provide decisions about significance that are relatively invariant over alternative choices of family size.

Table 8 presents mean *NAEP* mathematics scale scores (and standard errors) for the nation, for both 1990 and 1992 at three grades, and for various demographic subgroups that differ by Gender, Race/Ethnicity, Type of Community, and Region of the country (Mullis, Dossey, Owen, & Phillips, 1993, p. 18). Of the 45 contrasts between 1990 and 1992 results,

Table 7.

Mean (and standard error) 8th-grade mathematics achievement change by state, 1990 to 1992, t, p-value, and p_{crit}-values for four multiple comparison adjustments, $m = 34$ (df taken as 60).

State	$\bar{X}_{92}-\bar{X}_{90}$ (se)	t	(p-value)	p_{UNA}	$p_{FDR}(i)$	$p_{HOC}(i)$	p_{BON}
GA	-0.323 (1.77571)	-0.18190	(.42814)	.025	.025000	.025000	.000735
AR	-0.777 (1.48529)	-0.52313	(.30141)	.025	.024265	.012500	.000735
AL	-1.568 (2.01745)	-0.77722	(.22004)	.025	.023529	.008333	.000735
NJ	1.565 (1.92728)	0.81203	(.20999)	.025	.022794	.006250	.000735
NE	1.334 (1.52772)	0.87320	(.19320)	.025	.022059	.005000	.000735
ND	1.526 (1.68552)	0.90536	(.18445)	.025	.021324	.004167	.000735
DE	1.374 (1.34651)	1.02042	(.15581)	.025	.020588	.003571	.000735
MI	2.215 (1.84727)	1.19906	(.11761)	.025	.019853	.003125	.000735
LA	2.637 (2.07943)	1.26814	(.10482)	.025	.019118	.002778	.000735
IN	2.149 (1.63556)	1.31392	(.09694)	.025	.018382	.002500	.000735
WI	2.801 (1.96269)	1.42713	(.07936)	.025	.017647	.002273	.000735
VA	2.859 (1.92992)	1.48141	(.07187)	.025	.016912	.002083	.000735
WV	2.331 (1.39639)	1.66930	(.05013)	.025	.016176	.001923	.000735
MD	3.399 (1.92320)	1.76737	(.04113)	.025	.015441	.001786	.000735
CA	3.777 (2.11460)	1.78615	(.03956)	.025	.014706	.001667	.000735
OH	3.466 (1.85022)	1.87329	(.03295)	.025	.013971	.001563	.000735
NY	4.893 (2.53195)	1.93250	(.02901)	.025	.013235	.001471	.000735
PA	4.303 (2.20545)	1.95108	(.02786)	.025	.012500	.001389	.000735
FL	3.784 (1.93266)	1.95792	(.02745)	.025	.011765	.001316	.000735
WY	2.226 (1.09641)	2.03026	(.02339)	.025 *	.011029	.001250	.000735
NM	2.334 (1.14816)	2.03282	(.02325)	.025 *	.010294	.001190	.000735
CT	3.204 (1.53443)	2.08807	(.02052)	.025 *	.009559	.001136	.000735
OK	4.181 (1.75467)	2.38278	(.01018)	.025 *	.008824	.001087	.000735
KY	4.327 (1.61804)	2.67422	(.00482)	.025 *	.008088 *	.001042	.000735
AZ	4.994 (1.85110)	2.69785	(.00452)	.025 *	.007353 *	.001000	.000735
ID	2.956 (1.06775)	2.76845	(.00374)	.025 *	.006618 *	.000962	.000735
TX	5.645 (1.88770)	2.99041	(.00202)	.025 *	.005882 *	.000926	.000735
CO	4.326 (1.38868)	3.11519	(.00141)	.025 *	.005147 *	.000893	.000735
IA	4.811 (1.48805)	3.23309	(.00100)	.025 *	.004412 *	.000862	.000735
NH	4.422 (1.35399)	3.26591	(.00090)	.025 *	.003676 *	.000833	.000735
NC	7.265 (1.58701)	4.57779	(.00001)	.025 *	.002941 *	.000806 *	.000735 *
HI	5.550 (1.17134)	4.73817	(.00001)	.025 *	.002206 *	.000781 *	.000735 *
MN	6.421 (1.35226)	4.74836	(.00001)	.025 *	.001471 *	.000758 *	.000735 *
RI	5.097 (0.94844)	5.37407	(.00000)	.025 *	.000735 *	.000735 *	.000735 *
t_{crit}				2.00	2.47	3.30	3.33

* Confident direction of change.

Table 8.

From the *NAEP 1992 Mathematics Report Card for the Nation and the States* (Mullis, Dossey, Owen, & Phillips, 1993, p. 18).

TABLE 5 Average Mathematics Proficiency by Gender, Race/Ethnicity, Type of Community, and Region

	Assessment Years	Grade 4	Grade 8	Grade 12
Male	1992	220(0.8)>	267(1.1)>	301(1.1)>
	1990	214(1.2)	263(1.6)	297(1.4)
Female	1992	217(1.0)>	268(1.0)>	297(1.0)>
	1990	212(1.1)	262(1.3)	292(1.3)
White	1992	227(0.9)>	277(1.0)>	305(0.9)>
	1990	220(1.1)	270(1.4)	300(1.2)
Black	1992	192(1.3)	237(1.4)	275(1.7)>
	1990	189(1.8)	238(2.7)	268(1.9)
Hispanic	1992	201(1.4)	246(1.2)	283(1.8)>
	1990	198(2.0)	244(2.8)	276(2.8)
Asian/Pacific Islander	1992	231(2.4)	288(5.5)	315(3.5)
	1990	228(3.5)	279(4.8)!	311(5.2)
American Indian	1992	209(3.2)	254(2.8)	281(9.0)
	1990	208(3.9)	246(9.4)	288(10.2)!
Advantaged Urban	1992	237(2.1)	288(3.6)	316(2.6)
	1990	231(3.0)	280(3.2)	306(6.2)
Disadvantaged Urban	1992	193(2.8)	238(2.6)<	279(2.4)
	1990	195(3.0)	249(3.8)!	276(6.0)
Extreme Rural	1992	216(3.6)	267(4.6)	293(1.9)
	1990	214(4.9)	257(4.4)	293(3.3)
Other	1992	219(0.9)>	268(1.1)>	300(0.9)>
	1990	213(1.1)	262(1.7)	295(1.3)
Northeast	1992	223(2.0)>	269(2.7)	302(1.5)
	1990	215(2.9)	270(2.8)	300(2.3)
Southeast	1992	210(1.6)>	260(1.4)	291(1.4)>
	1990	205(2.1)	255(2.5)	284(2.2)
Central	1992	223(1.9)>	274(1.9)>	303(1.8)
	1990	216(1.7)	266(2.3)	297(2.6)
West	1992	218(1.5)	268(2.0)>	298(1.7)
	1990	216(2.4)	261(2.6)	294(2.6)

>The value for 1992 was significantly higher than the value for 1990 at about the 95 percent confidence level. < The value for 1992 was significantly lower than the value for 1990 at about the 95 percent confidence level. ! Interpret with caution – the nature of the sample does not allow accurate determination of the variability of this estimated statistic. The standard errors of the estimated proficiencies appear in parentheses. It can be said with 95 percent confidence for each population of interest, the value for the whole population is within plus or minus two standard errors of the estimate for the sample. In comparing two estimates, one must use the standard error of the difference (see Appendix for details).

Mullis et al. reported 21 to represent significant change at $\alpha = .05$, without adjustment for multiplicity.[6] To reduce the number of erroneous conclusions, as well as to control the probability of such errors, a multiple comparison procedure *should* be employed. Defining the family of comparisons conservatively as all tests in Table 8, $m = 45$. Using either the Bonferroni adjustment or the Hochberg procedure with $m = 45$, we find 7 confident directions of change; the FDR procedure yields 14 confident directions of change.

It might be argued that three separate families of comparisons are involved here, one family for each grade level, each with $m = 15$. Based on that approach, the Bonferroni adjustment and the Hochberg procedure both yield 10 confident directions; the FDR procedure results in the same 14 confident directions as before.

As still another alternative, it might be decided that there are four families of comparisons, one for each background variable presented in the table: Gender, Race/Ethnicity, Type of Community, and Region. For Gender, there are $m = 6$ contrasts (for 1990 to 1992 change for Females and Males at each of three grade levels); for Race/Ethnicity, there are $m = 15$ tests (for change for each of five Race/Ethnicity groups — White, Black, Hispanic, Asian/Pacific Islander, and American Indian — at each grade level); for Type of Community, there are $m = 12$ tests (for four community types — Advantaged Urban, Disadvantaged Urban, Extreme Rural, and Other — at each grade); and for Region, there are $m = 12$ tests (for four Regions — Northeast, Southeast, Central, and West — at each grade level). Accumulating results for each of these four families over all 45 comparisons (6+15+12+12), the Bonferroni procedure results in the same 10 confident directions as above, the Hochberg procedure results in 12 confident directions, and the FDR procedure results in 16 confident directions (14 as before with 2 additional).

As anticipated, the FDR procedure was the most consistent multiplicity adjustment across the four different definitions of family: With one large family of $m = 45$ or with three families corresponding to student's grade level ($m = 15$), the same 14 confident directions appeared using the FDR. When the 45 comparisons were divided into four families ($m = 6$, 15, 12, and 12), the FDR produced two additional confident directions. The Hochberg and Bonferroni procedures were more conservative as well as less consistent. The Bonferroni adjustment produced 7, 10, and 10 confident directions, and the Hochberg procedure produced 7, 10, and 12 confident directions for the three alternative family definitions. Of course, when no adjustment is made there are no inconsistencies — whatever the family size — because $p_{crit} = .05$ is applied throughout.

Even with the largest family size, $m = 45$, the FDR procedure admitted more confident directions than the number provided by the conventional Bonferroni applied to the smaller — and more lenient — family definitions. This suggests, once again, an increase in statistical power from the use of the FDR.

Applications to Simulated Data

When applying adjustment procedures to real data, we can count how many

differences are discovered to have a confident direction, but we cannot know how many of these are erroneous. We can, as we have done, compare the statistical power of various procedures anecdotally in terms of their performance on observed differences (population values *plus* specific random perturbations). But, to learn about erroneous discovery, in particular about false discovery rates, we must turn to some form of simulation.

We have chosen, mainly for simplicity, to simulate two sets of analyses for 48 "states," where perturbed values from a simple configuration (variously scaled) of 48 population values are either (a) compared to a fixed national value (m = 48 "uncorrelated differences") or (b) compared among themselves as differences of all possible pairs (m = 1128 "pairwise differences," with about 4% of pairs of pairs correlated). The simple configuration used is that of an idealized sample from a normal (Gaussian) distribution. We have assumed "large samples," taking the degrees of freedom for s^2 as infinite.

The simulated data are structured to be similar to the data from the *NAEP TSA*. For each of 48 states, mean "achievement levels," μ_i, are defined to be the approximate median values of each of 48 ordered random observations from a normal $(0,\sigma_A)$ distribution (for which s^2 = .98). Five conditions of effect size are studied.[7] For the "perinull" condition of negligible differences among the μ_i, the value of σ_A is set to 0.001; four non-null conditions are considered, with σ_A set to values of 0.3, 1.0, 3.0, and 5.0, respectively. In each case, for each of 10,000 replicates, an observed mean for each state, $\bar{X}_i$, is generated by adding a number randomly selected from a normal (0,1) distribution to the corresponding μ_i.

In the first of two families studied, each $\bar{X}_i$ is compared to a "national mean," treated here as a known constant, M. This results in m = 48 uncorrelated comparisons about which we wish to establish confidence about the sign of $\mu_i - M$. The second family is comprised of all pairwise comparisons where each $\bar{X}_i$ is compared with each $\bar{X}_j$, resulting in m = 1128 comparisons about which we wish to establish confidence about the sign of $\mu_i - \mu_j$.

The numbers of confident conclusions for the FDR technique and Hochberg's adjustment technique are compared with those resulting from conventional Bonferroni-adjusted p-values, in all cases with α = .05.

Familywise Error

Figure 6 presents plots of the familywise error against effect size for the FDR technique, the Hochberg technique, and the Bonferroni technique. The upper plot is for the family of 48 uncorrelated comparisons of state means with a constant national mean, M. The lower plot is for the family of all m = 1128 pairwise comparisons. In both cases, error rates are based on 10,000 replications.

In our perinull conditions for both uncorrelated-differences and pairwise-differences families, the FDR, the Hochberg, and the Bonferroni techniques maintain the familywise error rate at approximately $\alpha/2$ or below, as expected. (With negligible differences among the $\bar{X}_i$, claims of confident direction occur about $100\alpha\%$ of the time, and half of these claims are in the incorrect direction.)

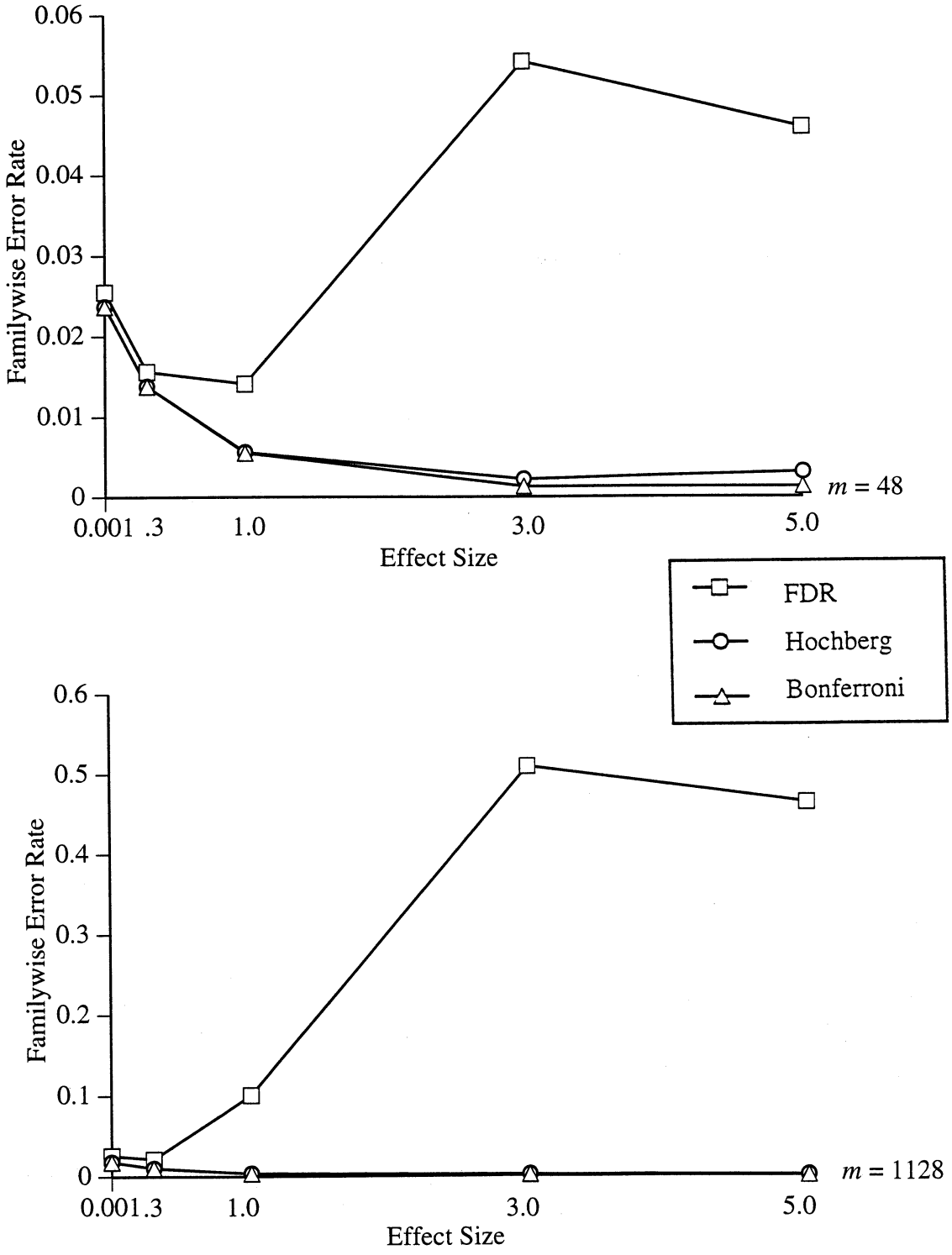

Figure 6.

Familywise error rates for the FDR, Hochberg, and Bonferroni techniques, 48 uncorrelated differences (above) and 1128 pairwise differences among the 48 (below).

For all three techniques, the familywise error rate is less than $\alpha/2$ in the small (σ_A = 0.3) and moderate (σ_A = 1.0) effect-size conditions for the uncorrelated-differences family, and in the small (σ_A = 0.3) effect conditions for the pairwise-differences family. Only in the large (σ_A = 3.0 and σ_A = 5.0) effect-size conditions for the uncorrelated-differences family, and in the moderate (σ_A = 1.0) and large ($\sigma_A \geq 3.0$) effect conditions for the pairwise-differences family, does the FDR technique fail to maintain the familywise error rate. The Hochberg and Bonferroni techniques maintain this error rate at or below $\alpha/2$ throughout.

False Discovery Rate

Figure 7 presents plots of the false discovery rates against effect size for the FDR, Hochberg, and Bonferroni adjustment techniques. Under all effect sizes, each adjustment procedure maintains a false discovery rate at or below $\alpha/2$, for both uncorrelated and pairwise families. We confirm the finding of Benjamini, Hochberg, and Kling (1994), that the false discovery rate is close to its maximum, $\alpha/2$, in the case of negligible differences among the μ_i (σ_A = 0.001), where there are the fewest claims of confident direction, but proportionally more (up to 50%) in the wrong direction.

Power

Figure 8 presents plots of the statistical power against effect size for each of the three adjustment techniques. Power is defined as what Hochberg and Tamhane (1987) refer to as *all-pairs power*, the probability of claiming confident direction for all true differences among all pairs; it is calculated as the average proportion of confident directions claimed over the 10,000 replications.

Under all effect-size conditions, for both uncorrelated and pairwise families, the FDR technique results in greater power than that for the Hochberg or Bonferroni procedures. The relative advantage in power for the FDR technique is greatest for the large pairwise family and for large effect sizes. (The increase in power of the Hochberg technique over the Bonferroni becomes detectable only for large effect sizes, $\sigma_A \geq 3.0$.) These results are consistent with the findings of Benjamini and Hochberg (in press) and Benjamini, Hochberg, and Kling (1994).

Some Further Simulation Results

Two more sets of simulated data were studied to try to tease apart the effects on error rates and statistical power of the partial dependence of pairwise comparisons, on the one hand, and of family size, on the other. In one condition, each of 1128 values of $\bar{X}_i$ is compared to a fixed known constant, M, the "national mean," yielding a family of uncorrelated differences of the same size as the family of pairwise differences that was studied and reported on above. In a second condition, 10 state mean values are compared among themselves as differences of all possible pairs (m = 10×9/2 = 45), with a family size similar to that for the 48 uncorrelated differences above. The same five conditions of effect size are studied, σ_A = 0.001, 0.3, 1.0, 3.0, and 5.0, and α = .05, also as before. Results are based on 10,000 replications.

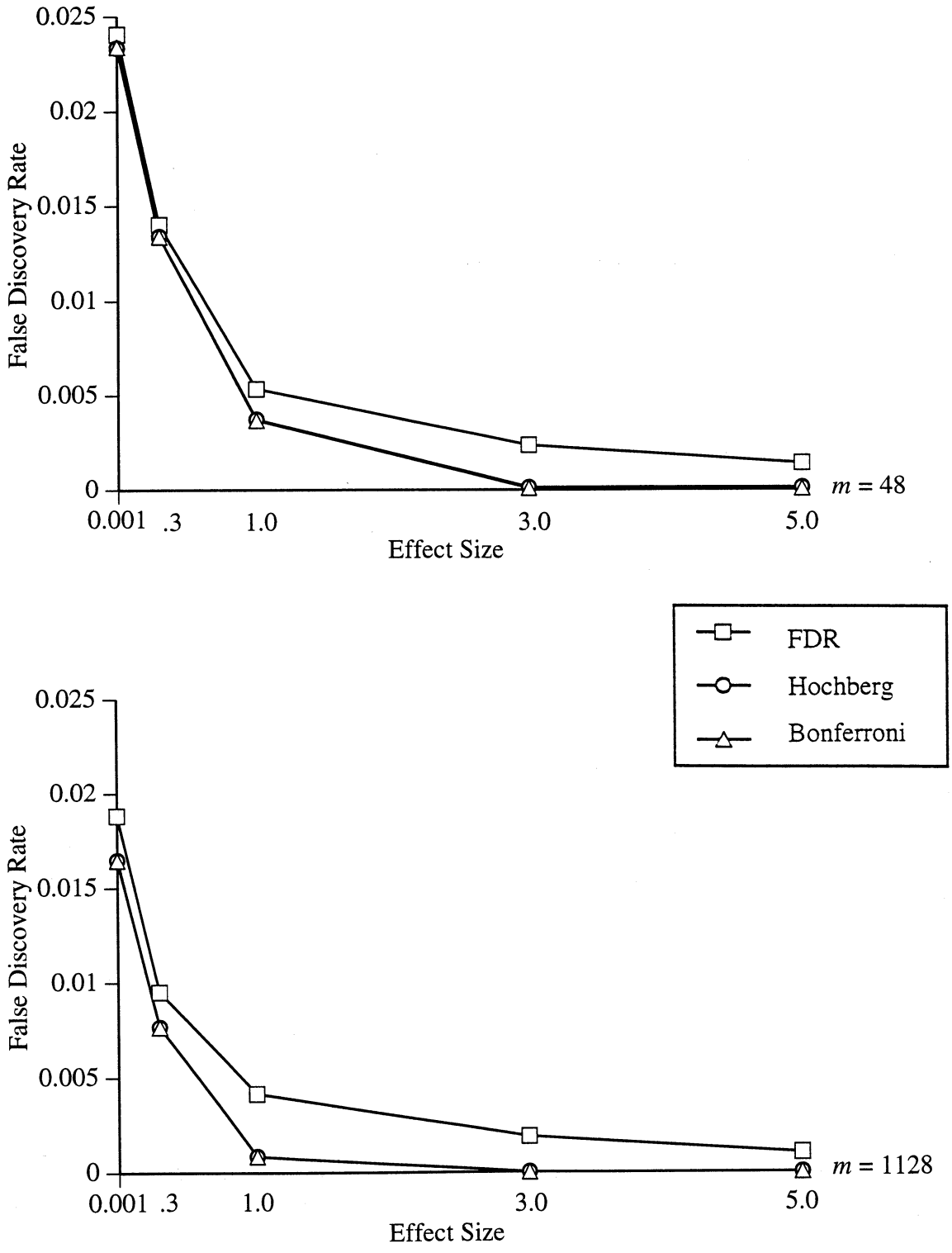

Figure 7.

False discovery rates for the FDR, Hochberg, and Bonferroni techniques, 48 uncorrelated differences (above) and 1128 pairwise differences among the 48 (below).

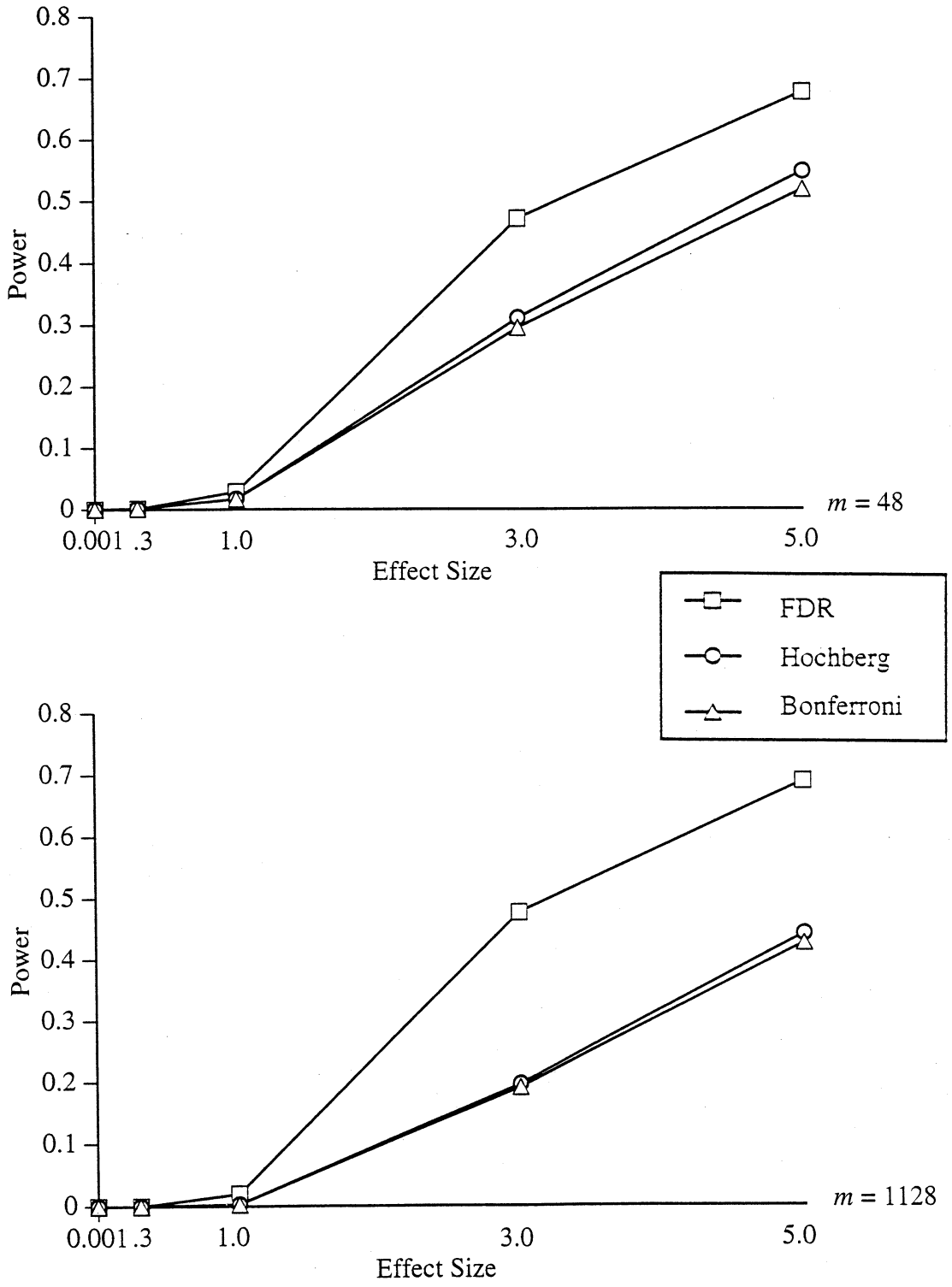

Figure 8.

Average statistical power for the FDR, Hochberg, and Bonferroni techniques, 48 uncorrelated differences (above) and 1128 pairwise differences among the 48 (below).

Figure 9 shows the familywise error for the FDR, the Hochberg, and the Bonferroni techniques for the five effect sizes. The upper plot is for the family of 1128 uncorrelated comparisons of state means with a constant national mean, and the lower plot is for the family of all $m = 45$ pairwise comparisons. The FDR technique does not maintain the familywise error rate at α in the large family of uncorrelated comparisons (upper plot); however, the FDR appears to provide ample protection against familywise error in the smaller family of all pairwise comparisons (lower plot). These results, compared with the familywise error rates displayed in Figure 6, suggest that the large family size rather than the nonindependence is driving the increased error rate.

The false discovery rates for each of the three adjustment techniques are shown in Figure 10. As in Figure 7, each adjustment maintains a false discovery rate at or below $\alpha/2$, for both the uncorrelated comparisons (upper plot) and the pairwise comparisons (lower plot) under all effect sizes.

Figure 11 presents plots of statistical power against effect size for the FDR, Hochberg, and Bonferroni adjustment techniques. The FDR technique results in greater power than the Hochberg or Bonferroni procedures under all effect-size conditions, for both the 1128 uncorrelated comparisons (upper plot) and the pairwise comparisons among 10 (lower plot). The relative advantage in power for the FDR technique is greatest for the large effect sizes. Comparing the results in Figure 11 with those presented in Figure 8, it is clear that the FDR advantage in power is associated with the large family size and is little affected by the dependence or the independence of the contrasts tested. (After all, at this family size, only about 8% of the pairs of paired comparisons are correlated.)

Conclusions

Results from the five analyses of the *Election Example* data and the seven analyses of the *NAEP TSA* data are summarized in Tables 9 and 10. The *recovery ratio* — (#FDR-confidences − #BON-confidences) / (#UNA-confidences − #BON-confidences) or how far the FDR moves from the Bonferroni toward the unadjusted rate — is plotted against the ratio (#BON-confidences+1)/m in Figure 12 for these examples. The figure shows a strong increasing trend in the recovery ratio with increases in the proportion of Bonferroni confidences; the gain in number of confident directions by the FDR procedure is greater when there are more confident directions by the Bonferroni, generally when family size is very large.

As expected, the change from the Bonferroni adjustment to the Hochberg procedure in total number of confident directions is small: 46 added to 1120. The effort involved is small, but the fact that the Bonferroni also generates matching confidence intervals, which the Hochberg procedure does not, is an important advantage of the Bonferroni technique.

The 493 confident directions added by the FDR technique to the 1166 from the Hochberg procedure are much more numerous and hence more valuable. If we are to give up the Bonferroni adjustment and dare not use an unadjusted approach, moving to the FDR technique seems an attractive choice.[8]

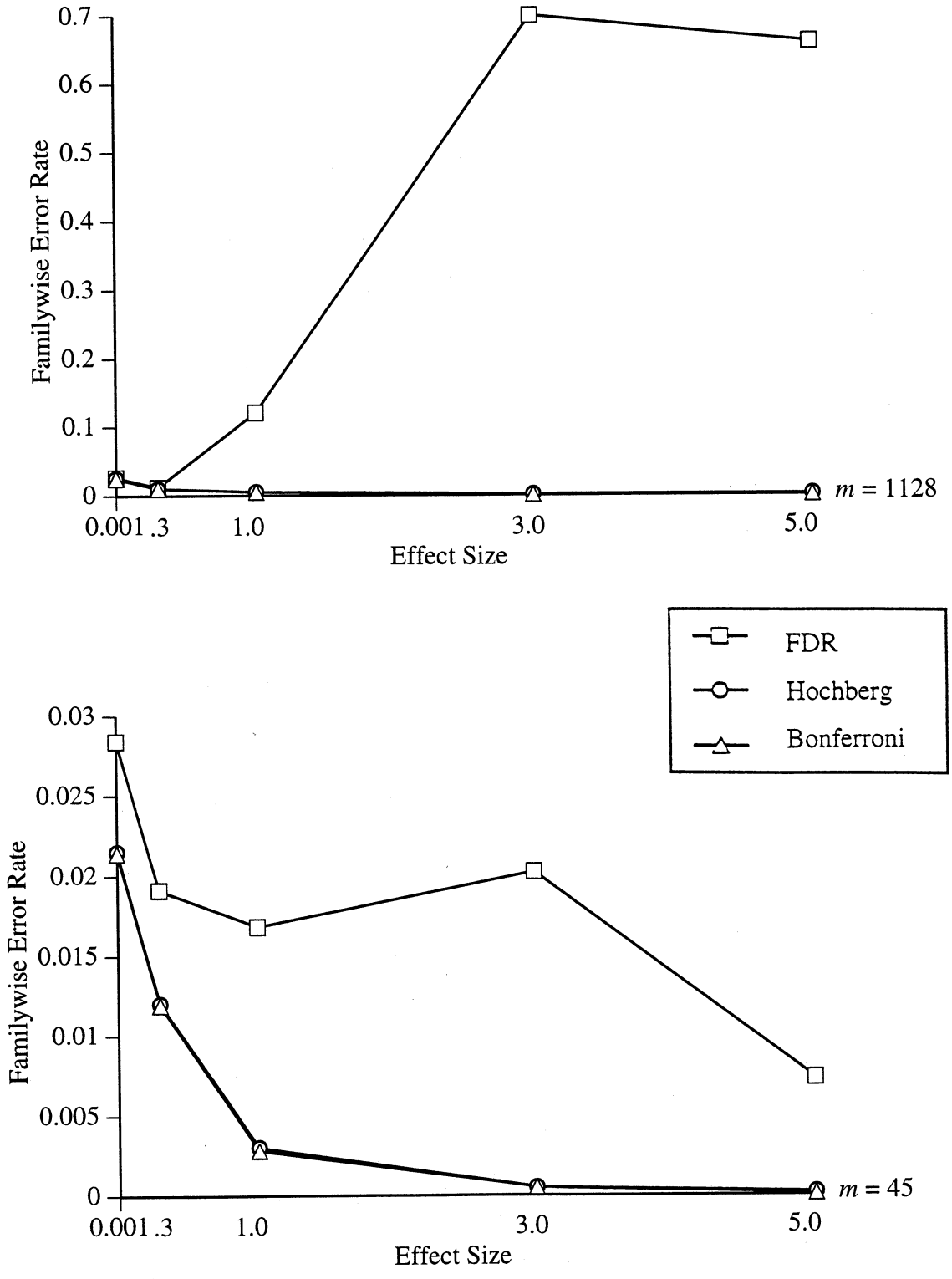

Figure 9.

Familywise error rates for the FDR, Hochberg, and Bonferroni techniques, 1128 uncorrelated differences (above) and 45 pairwise differences among 10 (below).

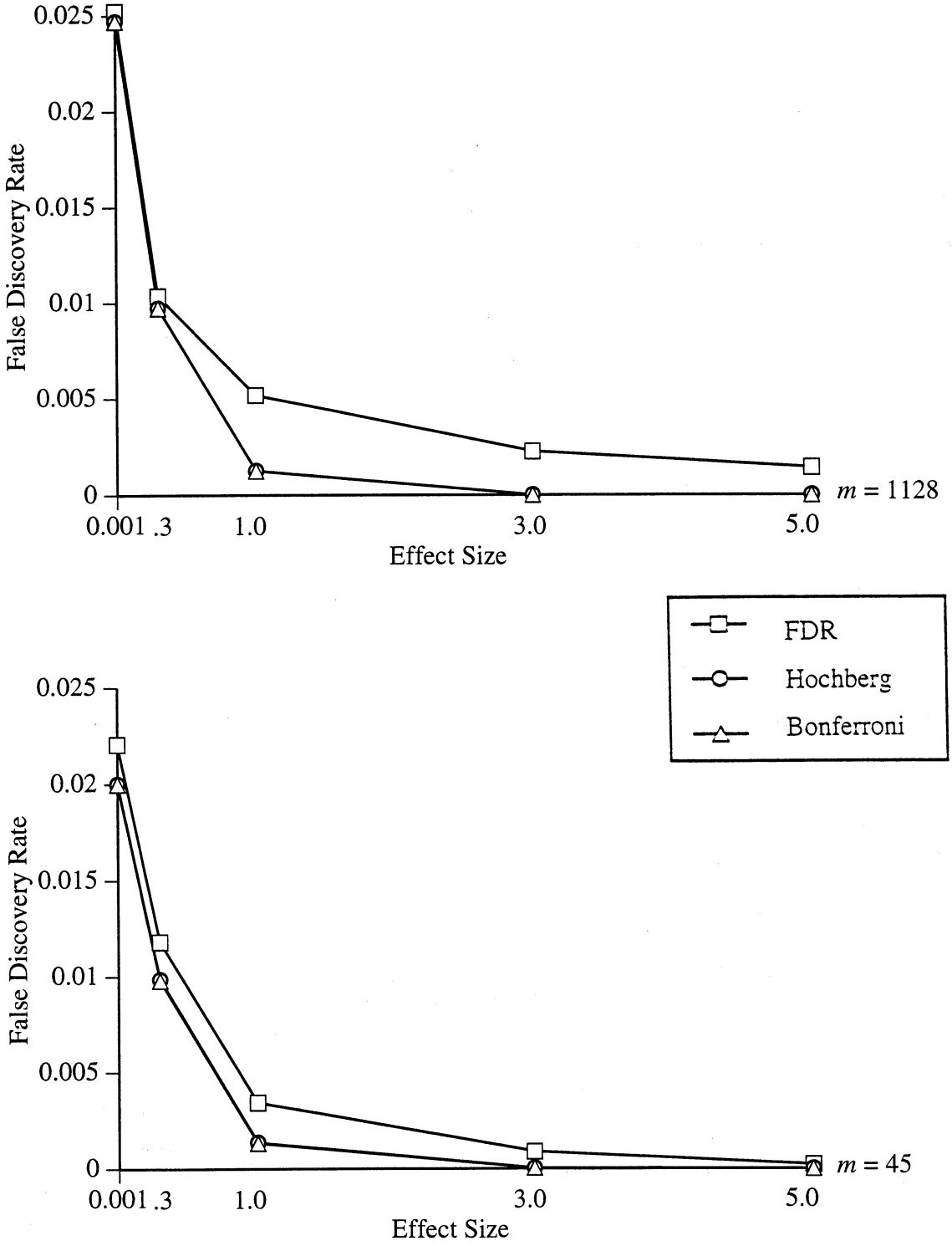

Figure 10.

False discovery rates for the FDR, Hochberg, and Bonferroni techniques, 1128 uncorrelated differences (above) and 45 pairwise differences among 10 (below).

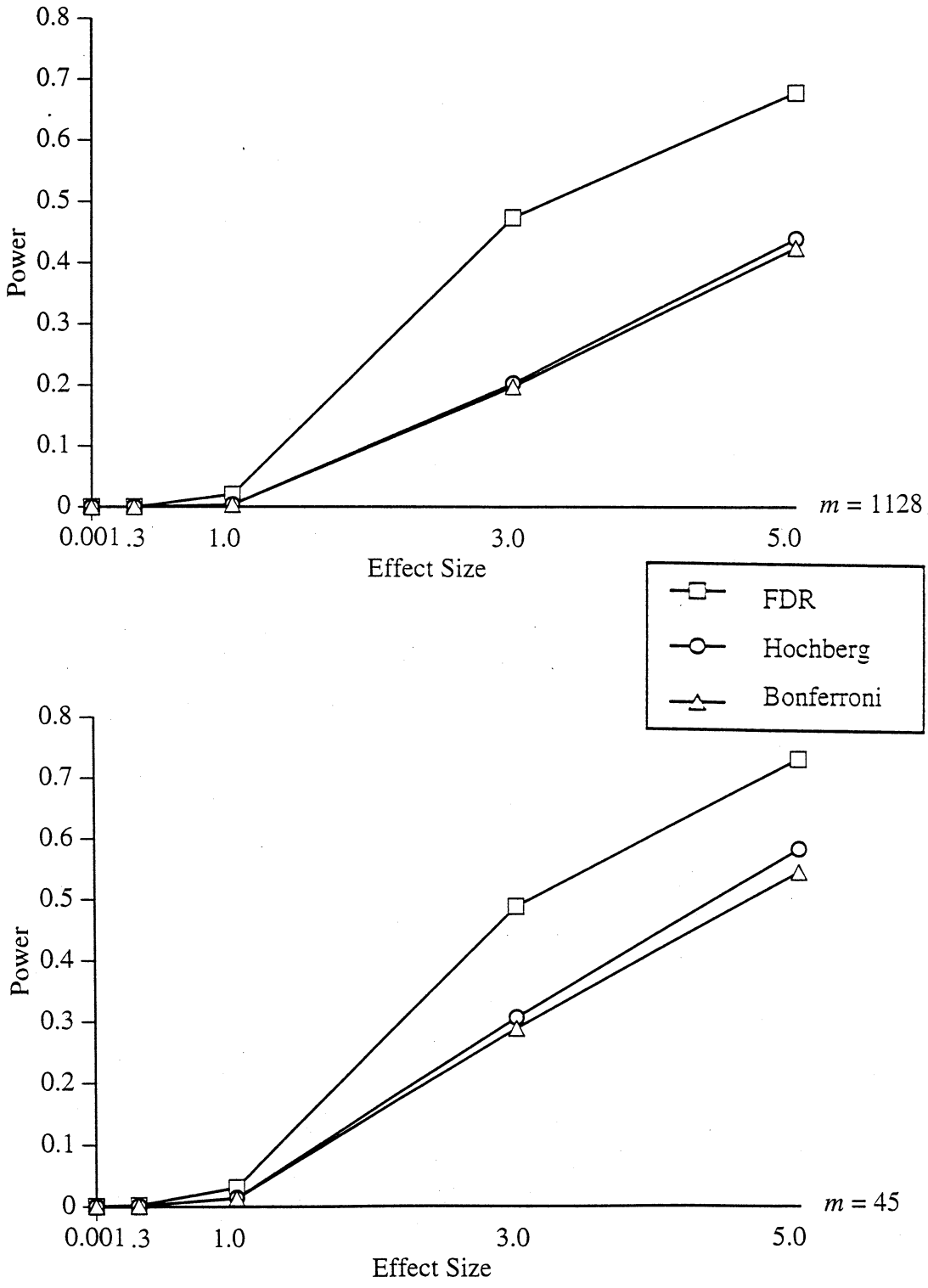

Figure 11.
Average statistical power for the FDR, Hochberg, and Bonferroni techniques, 1128 uncorrelated differences (above) and 45 pairwise differences among 10 (below).

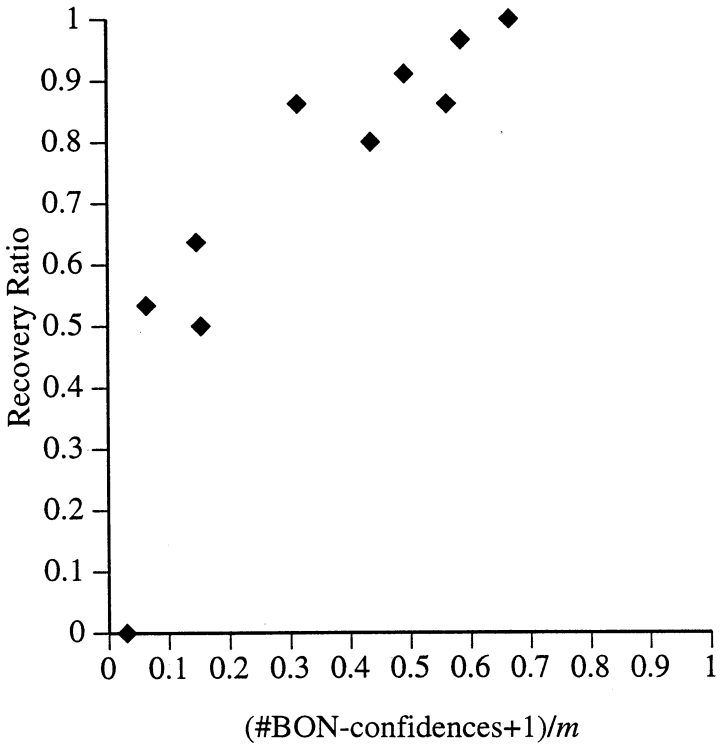

Figure 12.
Recovery ratio:
(#FDR-confidences – #BON-confidences) / (#UNA-confidences – #BON-confidences)
as a function of (#BON-confidences+1)/*m*.

Table 9.
Summary of the *Election* and *NAEP TSA Example*s: Number of confident directions observed and upper bounds (in italics) on the number expected by chance in the perinull situation.

	Unadjusted[a]		FDR[b]		Hochberg		Bonferroni		*m*
Election Example									
Election Year	4	*0.15*	4	*0.125*	4	*0.025*	4	*0.025*	6
Grouping	29	*1.95*	20	*0.525*	12	*0.025*	11	*0.025*	78
State(G)	21	*0.97*	20	*0.525*	17	*0.025*	16	*0.025*	39
State	391	*18.52*	369	*9.250*	238	*0.025*	232	*0.025*	741
E × G	99	*9.75*	64	*1.625*	25	*0.025*	24	*0.025*	390
NAEP TSA Example									
Region	5	*0.15*	5	*0.150*	5	*0.025*	3	*0.025*	6
State(R)	100	*3.20*	96	*2.425*	74	*0.025*	71	*0.025*	128
State	432	*14.02*	418	*10.475*	294	*0.025*	275	*0.025*	561
Y × R	0	*0.40*	0	*0.025*	0	*0.025*	0	*0.025*	16
Change	15	*0.85*	11	*0.300*	4	*0.025*	4	*0.025*	34
Differential Change	4	*0.85*	0	*0.025*	0	*0.025*	0	*0.025*	34
All Pairwise	658	*20.50*	652	*16.325*	493	*0.025*	480	*0.025*	820
Total	1758	*71.32*	1659	*41.775*	1166	*0.300*	1120	*0.300*	2853
Fraction[c]	24.6		39.71		3886.7		3733.3		

[a] The italicized number is $0.025m$.
[b] The italicized number incorporates the "+1" denominator modification discussed in Appendix A, and equals 0.025 *plus* 1/40th of the number of assertions.
[c] Fraction = (Number observed) / (Upper bound for number expected by pure chance).

The FDR procedure demonstrates a large gain in power over the simple Bonferroni adjustment. As indicated from the results shown both here and in Benjamini and Hochberg's (in press) simulation studies, the power advantage of the FDR procedure increases with the number of comparisons when the true differences remain about the same size: The loss of power with increasing m for the FDR technique is slower than the corresponding loss of power for the Hochberg and the Bonferroni adjustments. The conservatism of the Bonferroni and Hochberg procedures is due to the small p_{crit} required for strong protection against Type I error.

Because of the discontinuity in the number of erroneous detections at the exactly null situation (when all differences are exactly zero), perinull situations with trivially small differences have only half the error rate of the (unrealistic) null situation. Holding any form of error rate to $100\alpha\%$ in the unrealistic null situation requires holding the corresponding error rate for the more realistic perinull situations to half of this ratio, $50\alpha\%$.

Table 10.

Summary of Table 9, with a focus on *m* and on the recovery ratio, namely, (#FDR-confidences – #BON-confidences) / (#UNA-confidences – #BON-confidences).

Family	*m*	#UNA-confidences	#FDR-confidences	#BON confidences	Recovery ratio	(*)
All Pairwise[a]	820	658	652	480	97%	.59
State	741	391	369	232	86%	.31
State[a]	561	432	418	275	91%	.49
E × G	390	99	64	24	53%	.06
State(R)[a]	128	100	96	71	86%	.56
Remaining 7[b]	213	78	60	38	55%	.21
Total	(2853)	(1758)	(1659)	(1120)	(84%)	(.40)

(*) = (#BON-confidences+1)/*m*; in last line (Total #BON-confidences+7) / (Total *m*)

[a] From *NAEP TSA Example*

[b] Totalled for *Election Example* (Election Year, Grouping, State(G)) and for *NAEP TSA Example* (Year, Region, Change, Conditional Change)

Both the FDR and the Hochberg techniques are easily implemented. The sizable gain in power with the FDR procedure makes it an approach worth considering whenever it is acceptable to entertain the particular redefinition of α that the FDR procedure invokes: Where, under the unrealistic null situation, α becomes the average ratio of the number of erroneous declarations of confidence to the total number of declarations of confidence. Under the perinull situation, it then is expected that no more than $50\alpha\%$ of the total number of declarations of confident direction will be erroneous, i.e., will occur when the population comparisons would show a difference in the opposite direction.

Today, the choice among pure methods for practical work would seem to lie between the FDR procedure on the one hand and either the Bonferroni or Hochberg procedure on the other. Each of the three authors believes that the FDR procedure is the best choice.

** An alternative approach **

Some may find moving "whole hog" to FDR confidence about direction a rather drastic step. Since "not all confidences about direction were created equal!" we might prefer to go ahead with two or more cut-offs instead of one. A simple choice would be:

- if a direction is Bonferroni-confident, report it using the word "clearly";
- if a direction is FDR-confident but not Bonferroni-confident, report it without using any strengthening qualifier;
- otherwise, do not mention.

For the 78 differences between Groupings in the *Election Example* (see Figure 1 and Table 9), we would report "clearly confident" differences between Groupings for those 11 comparisons with the largest absolute values of t, we would claim 18 confident directions for those with intermediate values of t, and we would say nothing about the 49 remaining differences.

Another way to look at our examples is to ask, "How trustworthy are the additional confidences-in-direction statements discovered by the FDR?" After all, the Bonferroni-confidence statements are mostly free of error, so perhaps the additional supplementary confidences about directions are shot full of error!

To answer this question, we must first answer two other questions: (a) How many more comparisons are we FDR-confident about than we are Bonferroni-confident about? and, (b) What is the corresponding increase in erroneous statements (which ought to be at most the increase for pure chance for the various perinull situations)? Table 11 contains the numbers and the ratios. They indicate that when we look at only the FDR *supplementary* confidence statements, perhaps no more than 1 in 10 to 15 is likely to be erroneous.

Table 11.
Increases (FDR *vs* Bonferroni) in confidences and potentially incorrect confidences, and ratios, for examples.

Family	m	Increase in confidences	Increase in potential incorrect[a]	Ratio
All Pairwise[b]	820	172	16.3	10.6
State	741	137	9.2	14.9
State[b]	561	143	10.4	13.7
E × G	390	40	1.6	25.0
State(R)[b]	128	25	2.4	10.4
Remaining 7[c]	213	22	1.5	14.7
Total	2853	539	41.4	13.0

[a] Upper bound for average actual number of erroneous confidences, leading to lower bound for Ratio.
[b] From *NAEP TSA Example*
[c] Totalled for *Election Example* (Election Year, Grouping, State(G)) and for *NAEP TSA Example* (Year, Region, Change, Conditional Change)

Notes

[1] A useful modification of the Bonferroni choice makes use of the (one-sided) α point on the distribution of the Studentized range; this controls, at α, the chance of making one or more errors — if confidence intervals are stated for all simple comparisons or, *a fortiori*, if confidence statements are only made about directions. This chance is slightly smaller than the average number of errors, so that such a procedure need not control the average number to be $\leq \alpha$, but, if only directions are considered, is nearly sure to do so.

[2] The published results were insufficiently precise, rounded to only one decimal. We are grateful to Eugene G. Johnson for providing the more precise data, to two and three decimals for means and standard errors, respectively, that are analyzed here.

[3] 128 = (Central: 8×7/2)+(NE: 8×7/2)+(SE: 9×8/2)+(West: 9×8/2).

[4] If we were to be consistent about our emphasis on comparisons and directionality, we would eschew the *F*-test and work with the (much more interpretable) Studentized range instead.

[5] For results in Tables 6 and 7, the number of students sampled within states is generally close to 2000; however, because of the clustered nature of the sample design and the use of plausible values in *NAEP*, the effective sample size per state is estimated to be about 30, so that the degrees of freedom for a pairwise mean comparison is about 60.

[6] In our attempt to replicate the findings of Mullis, Dossey, Owen, and Phillips (1993), we find only 20 confident directions, most likely because we used the (rounded) published values for means and standard errors, rather than the more precise values used for computation by the original investigators.

[7] "Effect size" is approximately 1.02 times the standard deviation of population values.

[8] The properties of the confidence interval from FDR deserve study. We will return to them in another report.

References

Benjamini, Y., & Hochberg, Y. (in press). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B.*

Benjamini, Y., Hochberg, Y., & Kling, Y. (1994). False discovery rate controlling procedures for pairwise comparisons. Unpublished manuscript.

Diaconis, P. (1985). Theories of data analysis: From magical thinking through classical statistics. In Hoaglin, D. C., Mosteller, F., & Tukey, J. W., *Exploring data tables, trends, and shapes*. New York: Wiley.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800-803.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Johnson, E. G., Mazzeo, J., & Kline, D. L. (1993). *Technical report of the NAEP 1992 Trial State Assessment program in mathematics*. Report 23-ST05. Washington, DC: National Center for Education Statistics.

Mullis, I., Dossey, J., Owen, E., & Phillips, G. (1993). *NAEP 1992 mathematics report card for the nation and the states*. Washington, DC: National Center for Education Statistics.

Shaffer, J. P. (1994). *Multiple hypothesis testing: A review*. Report 23. Research Triangle Park, NC: National Institute of Statistical Sciences.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Tukey, J. W. (1993). Where should multiple comparisons go next? In Fred M. Hoppe (Ed.), *Multiple comparisons, selection, and applications in biometry* (pp. 187-207). New York: Marcel Dekker, Inc.

Tukey, J. W., & Hoaglin, D. C. (1991). Qualitative and quantitative inference. In Hoaglin, D. C., Mosteller, F., & Tukey, J. W., *Fundamentals of exploratory analysis of variance*. New York: Wiley.

Tukey, J. W., Mosteller, F., & Hoaglin, D. C. (1991). Concepts and examples in analysis of variance. In Hoaglin, D. C., Mosteller, F., & Tukey, J. W., *Fundamentals of exploratory analysis of variance*. New York: Wiley.

Appendix A
Null Behavior of Benjamini and Hochberg's FDR procedure

Any "false discovery rate" procedure must squirm a little at an exact null hypothesis, where all discoveries are by definition false. In the Benjamini and Hochberg (in press) formulation, this squirming is done by controlling the average, over realizations, of

false discoveries / # discoveries (if there are any discoveries)

or

0 (if there are no discoveries),

and requiring that average to be less than the nominal α.

Another, often more palatable, squirm is to ask about the application of both the Bonferroni and FDR adjustments to the same data set. Even if we take the more classical view, and consider an exact null hypothesis against all possible alternatives ("= 0" *vs* "≠ 0"), we easily find some illuminating results.

First, note that finding no FDR significances implies finding no Bonferroni significances. At the exact null hypothesis, then, any realization with no errors by the FDR procedure has no errors by the Bonferroni.

Second, in the case of independence, we can easily calculate the probabilities, P, of exact outcomes, namely the probabilities (neglecting terms of higher order in α and terms involving $1/m$) of zero, one, or two errors of any kind (three or more errors are negligibly likely):

- no error by either Bonferroni or FDR: $P = [1-(\alpha/m)]^m \approx e^{-\alpha} \approx 1-\alpha$,
- exactly one Bonferroni error: $P = m(\alpha/m)[1-(\alpha/m)]^{m-1} \approx \alpha e^{-\alpha} \approx \alpha$,

If exactly two p_i are less than $2\alpha/m$, an event of null-hypothesis probability

$$[m(m-1)/2](2\alpha/m)^2[1-(2\alpha/m)]^{m-2} < 2m(m-1)(\alpha/m)^2 < 2\alpha^2 ,$$

then there are at least two FDR events. Of these two "less-than-$2\alpha/m$" situations, ¼ have no Bonferroni errors (because both $p_i > \alpha/m$), ½ have one Bonferroni error, and ¼ have two Bonferroni errors. Accordingly:

- exactly two FDR errors, both Bonferroni: $P < \alpha^2/2$
- exactly two FDR errors, one Bonferroni: $P < \alpha^2$
- exactly two FDR errors, none Bonferroni: $P < \alpha^2/2$

(In fact, all the ">" differ from "=" by terms in α^3, α^3/m, and higher.) For $\alpha = .05 = 5\%$, the occurrence probabilities are approximately:

- no error: $P \approx 95\%$,
- one Bonferroni error (and no other FDR error): $P \approx 5\%$,
- two Bonferroni errors: $P \approx (1/8)\%$,
- one Bonferroni error (also FDR) and another FDR error: $P \approx (1/4)\%$, and
- two FDR errors and no Bonferroni errors: $P \approx (1/8)\%$.

The ratio of

$$P(\text{exactly two FDR errors})/P(\text{one Bonferroni error}) < 2\alpha^2/\alpha = 2\alpha$$

shows that the increase in errors at the null hypothesis from using the FDR procedure is only a small fraction of the errors already made using the Bonferroni (1/10 where $\alpha = 5\%$, nondirectional; 1/20 where $\alpha = 5\%$, directional). In other words, the required squirm is trivial and not worth any serious concern at all.

The increase in frequency of error is negligible, but the raw ratio:

# of false assertions / # of assertions

is not well controlled; in fact, this ratio is arbitrarily close to 50% when the true differences are close enough to zero — not only for the FDR, but for all procedures, including the very conservative ones: Bonferroni, Hochberg's (1988), the Studentized range, or what have you. This is true because the two signs, + and −, are nearly equally likely for any comparison where the true difference is very small.

If we want to be precise, we need to make some kind of modification. Benjamini and Hochberg (in press) showed that if two modifications were made, their method controls the (modified) FDR. The two modifications are:

- average the ratios of false assertions to all assertions, one ratio for each realization,
- when no assertion is made — so that this ratio is 0/0 — assign zero.

Another modification which is simpler to describe to users is to add 1 to the denominator of the raw ratio for each family, making it

false discovery proportion = # of false assertions / (# of assertions + # of families) .

We have marked this indicator by using "proportion" instead of "rate."

For R realizations, and with all the differences close to zero, the numerator will be, nearly

$$R(\alpha/2+2\alpha^2/4+...) ,$$

while the denominator will be, nearly

$$R(\alpha+1) ,$$

with ratio nearly

$$(\alpha/2+\alpha^2/2)/(\alpha+1) = \alpha/2 .$$

If this modification works well enough, users can honestly be told that we are trying to control the number of false assertions at no more than the number permitted by Bonferroni *plus* one for every $2/\alpha$ assertions (every 40, if $\alpha = .05$).

For those who find it necessary to work with "$\neq 0$" instead of with direction, the raw ratio is close to 100% (instead of 50%) and clearly requires modification at least as seriously as when directions are used.

Appendix B
Structure of Heteroscedastic Analyses of Simple, Trend-free, and Step-free Interaction

In the *Election Example*, consider the four observed values — y_{32}, y_{36}, y_{40}, and y_{44} — and the means across years, $\bar{y} = (y_{32}+y_{36}+y_{40}+y_{44})/4$ for (i) a specific State, (ii) a Grouping mean, (iii) a deviation of a State from its Grouping mean, (iv) an overall mean, or (v) the deviation of a Grouping mean from the overall mean. We can take a first step toward isolating and assessing some kind of interaction in several ways.

Simple interaction I. Most simply, we can form $z_1 = y_{32}-\bar{y}$, $z_2 = y_{36}-\bar{y}$, $z_3 = y_{40}-\bar{y}$, and $z_4 = y_{44}-\bar{y}$, and work with these deviations. In such an analysis, we presumably have not done anything to filter out trends. As a result, z_1 and z_4 are likely to be more variable than z_2 and z_3 because they include larger contributions from trends. Moreover, if the trends do not follow straight lines, either z_1 and z_4 might be more variable than the other. Accordingly, it may be desirable for our analysis to allow for such differences in variance. For the analysis of the interaction as differences within Grouping for each Election Year (applied to deviations of Election Year from overall), we have numbers for 39 States in 13 Groupings, and can estimate a variance on 26 = 39−13 degrees of freedom and can work with the 78 = (13×12)/2 simple comparisons between Groupings (within an Election Year). Doing this for each Election provides 312 = 4×78 comparisons as before — the novelty being the appearance of different variances in the denominators of the Student's *t*-values.

Simple interaction II. Going the other way, we have six kinds of Election comparisons — $z_1 = y_{36}-y_{32}$, $z_2 = y_{40}-y_{32}$, $z_3 = y_{44}-y_{32}$, $z_4 = y_{40}-y_{36}$, $z_5 = y_{44}-y_{36}$, and $z_6 = y_{44}-y_{40}$ — each of which can reasonably have its own variance. Each of these six has values for 39 States in 13 Groupings, so that again we can have separate estimates of variability on 26 = 39−13 degrees of freedom. And we can use these estimates to evaluate 13 *t*-values for each of the six Election comparisons, looking at:

Grouping mean *minus* grand mean (13 values)

against a background of:

State value *minus* Grouping mean (39 values, 26 *df*).

Again, we have 78 = 13×6 *t*-values, as we had before, the difference again being a diverse set of variances, one for each of the six kinds of comparisons between Elections.

Trend-free interaction. Another approach is to remove a linear trend from y_{32}, y_{36}, y_{40}, and y_{44}, leaving as the four residuals

$$z_1 = (3y_{32}-4y_{36}-y_{40}+2y_{44})/10 ,$$
$$z_2 = (-4y_{32}+7y_{36}-2y_{40}-y_{44})/10 ,$$
$$z_3 = (-y_{32}-2y_{36}+7y_{40}-4y_{44})/10 , \text{ and}$$
$$z_4 = (2y_{32}-y_{36}-4y_{40}+3y_{44})/10 ,$$

each of which will surely deserve separately estimated variances. We can treat z_1, z_2, z_3 and z_4 just as we treated those above.

Step-free interaction. Alternatively, we might take away not a linear trend, but a step difference between the two earlier Elections and the two later Elections working with

$$z_1 = (y_{32}-y_{36})/2 \text{ and } -z_1 = (y_{36}-y_{32})/2 \text{ , and}$$
$$z_2 = (y_{40}-y_{44})/2 \text{ and } -z_2 = (y_{44}-y_{40})/2 \text{ .}$$

Because the two degrees of freedom left after fitting both a mean and a step to each set of the four values can be so simply described by simple comparisons, we get two values to carry on instead of four, and two sets of 78 Grouping comparisons within Election comparison instead of four. Again the separate estimates of variability, based on the State value (of an x_i) *minus* the Grouping value (based on the same x_i), will have 26 degrees of freedom each.

Trend and Step. To go a little further with each of the last two analyses, we can analyze what is taken out as well as what is left after taking out. For the trend elimination, we can work with the estimate of slope

$$z_3 = (-3y_{32}-y_{36}+y_{40}+3y_{44})/20$$

and in the step-free analysis we can work with the estimate of step

$$z_4 = (y_{32}+y_{36}-y_{40}-y_{44})/2$$

in each case finding a new estimate of variance, based on 26 degrees of freedom and looking at $78 = (13\times12)/2$ differences of z_3 or z_4.

Table B-1 displays summary results for the four adjustment techniques applied to the heteroscedastic analyses of the E × G interaction effects. Of the 1014 total comparisons in six different families, the unadjusted per comparison approach results in 381 confident directions and the FDR technique results in 243 differences with confident direction. Applying strong control of the familywise error rate, the Hochberg technique results in 73 confident directions of which the Bonferroni adjustment results in the detection of 70 differences with confident direction.

Table B-1.
Number of comparisons from the heteroscedastic analyses of E × G Interaction Effects (from the *Election Example*) with confident direction.

	m	Unadjusted	FDR	Hochberg	Bonferroni
Simple interaction I	312	140	101	31	30
Simple interaction II	78	29	19	7	7
Trend-free	312	112	73	12	11
Step-free	156	34	0	0	0
Trend	78	31	23	9	9
Step	78	35	27	14	13

Appendix C
Comparison of Critical *t*-values

Some further insight into the performance of the three alternative adjustment approaches can be obtained from a table of the critical values of t (rather than the critical values of p). Indeed, we can learn even more from the critical values of t^2 and the ratios

$$\text{adjusted-}t_{crit}^2 \text{ / unadjusted-}t_{crit}^2$$

which, because observed t^2 is proportional to sample size, is an indication of the factor by which sample size must be increased to compensate for a given standard of multiplicity protection. Tables C-1 and C-2 show the numerical results for 2 of the 11 examples, Main Effects for Election Year from the *Election Example* and All Pairwise Comparisons among States from the *NAEP TSA Example*. (Recall that the FDR and Hochberg procedures work sequentially from the least significant comparison, the largest i, to the most significant, $i = 1$, as illustrated in Table C-1.) Also shown in the tables are the cutoff points for confident direction for the FDR and Hochberg procedures.

Table C-1.
Ratios of adjusted-t_{crit}^2 to unadjusted-t_{crit}^2 for Main Effects for Election Year.

i	FDR-t_{crit}^2	ratio	Hochberg-t_{crit}^2	ratio	Bonferroni-t_{crit}^2	ratio
6	4.1132	1.000	4.1132	1.000	7.79511	1.895
5	4.4617	1.085	5.4712	1.330	7.79511	1.895
4	4.8968	1.191	6.3053	1.533	7.79511	1.895
3	5.4712	1.330	6.9140	1.681	7.79511	1.895
2	6.3053	1.533	7.3956	1.798	7.79511	1.895
1	7.7951	1.895	7.7951	1.895	7.79511	1.895

Table C-2.
Ratios of adjusted-t_{crit}^2 to unadjusted-t_{crit}^2 for All Pairwise Comparisons among States.

i	FDR-t_{crit}^2	ratio	Hochberg-t_{crit}^2	ratio
820	3.8415	1.000	3.8415	1.000
819	3.8435	1.001	5.0239	1.308
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
657	4.2153	1.097	13.0402	3.395
656	4.2179	1.098	13.0516	3.398
655	4.2205	1.099	13.0629	3.401
654	4.2231	1.099	13.0741	3.403
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
509	4.6509	1.211	14.2477	3.709
508	4.6543	1.212	14.2537	3.711
507	4.6577	1.212	14.2597	3.712
506	4.6611	1.213	14.2657	3.714
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2	14.7623	3.843	16.0698	4.183
1	16.0721	4.184	16.0721	4.184

Note: Bonferroni-t_{crit}^2 and the ratio of Bonferroni-t_{crit}^2 to unadjusted-t_{crit}^2 are constant over i; both values are equal to those for FDR-t_{crit}^2 or Hochberg-t_{crit}^2 at $i = 1$.

Summarized in Table C-3 for all examples, are the ratios of the adjusted to unadjusted values of t_{crit}^2 for the smallest comparison confident in direction and the adjusted to unadjusted values of t_{crit}^2 for the largest comparison not confident in direction. For both the FDR and the Hochberg procedures, it can be seen that the discrepancy between the two ratios decreases with larger family sizes. For the ratio of t^2s for the smallest comparison with confident direction, the values for Bonferroni-confidence range from 1.895 to 4.475 and tend to increase with family size (when $m = 1$, the value of the ratio, adjusted-t_{crit}^2 / unadjusted-t_{crit}^2, is 1.0); the values for Hochberg-confidence have a similar range (1.335 to 4.249) and show a similar dependence on family size. However, the values of the FDR ratios, ranging from 1.086 to

2.633, show no substantial evidence of systematic change with changes in family size, if anything, drifting weakly in the opposite direction.

Table C-3.
Ratios of adjusted to unadjusted values of t_{crit}^2 for SMALLEST COMPARISON CONFIDENT IN DIRECTION (>) and LARGEST COMPARISON NOT CONFIDENT IN DIRECTION (<) for all examples ordered by m.

		Ratios for:				
	m	Bonferroni[a]	Hochberg[b]		FDR[b]	
			>	<	>	<
Main Effects for Election Year[c]	6	1.895	1.533	1.330	1.191	1.085
Main Effects for Region	6	1.913	1.335	1.000	1.086	1.000
Year × Region	16	2.174	——	2.174	——	2.174
State-by-State Change	34	2.633	2.557	2.540	1.466	1.429
State *vs* Average Change	34	2.633	——	2.633	——	2.633
States nested within Groupings	39	2.817	2.537	2.513	1.306	1.283
Main Effects for Grouping	78	3.430	3.332	3.322	1.672	1.647
States nested within Regions	128	3.233	2.783	2.774	1.109	1.105
Election Year × Grouping	390	4.100	4.063	4.062	1.857	1.849
States	561	4.087	3.649	3.647	1.112	1.111
States	741	4.475	4.249	4.248	1.320	1.319
Pairwise Comparisons of States[d]	820	4.184	3.712	3.711	1.099	1.098

[a] Values apply for all i, and are for possible comparisons, not just for those which occurred.
[b] When $i = 1$, ratios are the same as for Bonferroni; when $i = m$, ratios are 1.000.
[c] As shown in Table C-1.
[d] As shown in Table C-2.