The Annals of Statistics
2002, Vol. 30, No. 6, 1576–1594

JOHN W. TUKEY'S CONTRIBUTIONS TO MULTIPLE COMPARISONS

BY YOAV BENJAMINI AND HENRY BRAUN

Tel Aviv University and Educational Testing Service

This article provides a historical overview of the philosophical, theoretical and practical contributions made by John Tukey to the field of simultaneous inference. His early work, culminating in the monograph "The Problem of Multiple Comparisons," established him as one of the pioneers in the field, investing it with both academic respectability and a focus on practical problems. For many years afterward, Tukey only published sporadically in the area but remained convinced that multiplicity issues were of fundamental importance. During the last decade of his life, Tukey again devoted substantial attention to multiplicity, experimenting with different graphical representations of multiple comparison procedures and exploring the implications of new approaches to controlling family-wise error rates. He leaves a rich legacy that should engage and inspire statisticians for many years to come.

1. Introduction. The problem of multiplicity or simultaneous inference is concerned with making multiple inferences from a single set of data as well as the appropriate interpretation and use of the results. Though not always explicitly recognized, the problem occurs widely and often. The proper treatment of multiplicity, which should take into account the "tradeoff between extracting belief from data and payment of error" [Tukey (1991)], is regarded by many as a critical component in a disciplined program of scientific research.

While the issue of multiplicity was recognized by some investigators long ago, efforts to grapple with the problem were rather scattered until the post-WWII era. John Tukey is rightly considered one of the pioneers of the field. His early work, culminating with the informal release of the magisterial "The Problem of Multiple Comparisons" (PMC) in 1953, did much to shape the philosophy, mathematical development and practical applications of simultaneous inference.

Philosophy. Tukey proposed a proper framing of the problem of multiplicity, examined different kinds of error rates, their advantages and disadvantages, and considered the appropriate targets of inference as well as the quantification of the uncertainty surrounding those inferences.

Mathematical theory. Tukey developed techniques and probabilistic inequalities to prove properties of certain methods. He also suggested strategies for generating different methods for controlling error rates and provided heuristics for their operating characteristics.

Received September 2001; revised April 2002.

AMS 2000 subject classifications. 01-02, 62-03, 62F03.

Key words and phrases. John W. Tukey, simultaneous inference, mulitiplicity, multiple comparisons, false discovery rate.

Practical applications. Tukey developed a variety of methods, as well as a number of graphical techniques and reference tables to support their use. In addition he published interesting applications as exemplars of good statistical practice.

We claim, and attempt to substantiate in the following section, that not only was PMC the first full treatment of the subject, but also it largely set the terms for research and practice in the field for the next 35 years. Indeed there is hardly an aspect of the problem that has arisen over the years that was not addressed, often quite thoroughly, in the pages of PMC. No other work or writer has had comparable influence.

PMC summarizes Tukey's work in the field of multiple comparisons over the period 1947–1953. A long dry spell (in this area) followed as his attention shifted to other topics (many of which can be traced back to PMC), but he resumed intensive activity in this field during the last decade of his life. Throughout the second half of the twentieth century, the field of multiple comparisons has been a source of continuing debate at both the philosophical and methodological levels. Even when Tukey was less active in the area, he was strongly identified with a particular perspective on the topic and engaged in vigorous debate with other scholars (particularly David Duncan and Henry Scheffé).

In his second period of intense involvement, Tukey revisited the major philosophical issues, usually reconfirming his original views but with some elaboration based on the progress made since PMC appeared. He continued to suggest new methods while critiquing and applying various procedures developed by other investigators. Most importantly, perhaps, he set out an agenda for the future that identified the critical problems along with plausible directions for researchers and practitioners to explore. It is entirely possible that his later work, some of which has been published posthumously, will have a substantial impact on the field's future trajectory.

We will review in some detail John Tukey's contributions to this important area in theoretical and applied statistics. The structure of our presentation reflects the two separate periods of sustained activity. Section 2 highlights Tukey's contributions to simultaneous inference during the 1950s. Section 3 provides both a chronology of Tukey's later work and a more thematic discussion of his activities and influence on the field, spotlighting the 1990s. Some concluding remarks are offered in Section 4.

2. Early contributions.

2.1. *Introduction.* Our attention will be focussed on PMC, since Tukey's earlier work was fragmentary while the later work (from this period) expanded on specific issues. PMC was released in mimeograph form, and made available to others only by request. The fact that it went on to have such a powerful influence on the field is therefore both noteworthy and surprising. The interested reader is referred to Shaffer (1995) for further details or to Tukey [(PMC), pages 104–110]. It should be said at the outset that our claim for Tukey's contributions and influence is much greater than his.

> It is not the intention of the present account to claim or suggest that novel statistical ideas or components are here assembled into a coherent and workable approach to the multiple comparison problem and its relatives. The ideas and components are almost all old and well known, the novel feature is their assembly.
>
> (PMC), page 104

While this may be true in a narrow sense, the value of Tukey's work lies in the fact that his methodological innovations were embedded in a theoretical framework and a philosophy of practice that offered both the academic statistician and the practitioner a coherent approach to simultaneous inference.

2.2. *Philosophy.* While PMC does not provide a pithy characterization of the multiplicity problem, it illustrates how the problem arises in the context of scientific investigations. For example, Tukey notes that carrying out 250 independent tests of significance, each at the 0.05 level, will result on average in 12.5 apparently significant results when the intersection null hypothesis of no effects is true. Thus obtaining (say) 18 significant results is no cause for exultation [(PMC), pages 75–76]. This sort of argument, later dubbed "The Higher Criticism," found repeated use for pedagogical purposes.

Tukey believed that statisticians have much to learn from the activity of able scientists. The latter usually carry out exploratory analyses of their data, often desiring to attach some statistical meaning to their results. While remaining appreciative of serendipity, Tukey sought to discipline it in order to make it more valuable. He argued that it is important to distinguish between the two main sources of statistical problems: Administration or "for the moment" versus Science or "for the record" [(PMC), page 163]. Equally important is the recognition of the different principal aims of statistical activity, which he defined as action, indication and sanctification [(PMC), page 169]. Confidence intervals are identified with indication and significance tests with sanctification and action.

In an extended discussion of the pros and cons of confidence and significance, Tukey expresses a clear preference for confidence where it is feasible. In deciding among confidence procedures, he is comfortable using conventional statistical ideas to guide the choice [(PMC), pages 241–260]. Figure 9 on page 250, which illustrates the operating characteristics of three "pure" procedures and various combinations, anticipates the graphical representations employed extensively in the report on the Princeton Robustness Study [Andrews et al. (1972)].

Tukey repeatedly emphasized that intelligent control of multiplicity depends crucially on the appropriate choice of the family of statements, be it all determinations, all pairwise comparisons, all contrasts, all linear combinations, or the like. Different methods may prove useful for different purposes and one role for the statistician is to provide guidelines for use based on the nature of the problem and the aim of the analysis. Tukey goes on to establish some general criteria that multiplicity procedures ought to satisfy, providing a rationale for each. Later on, this framework is employed to attack the problem of developing new procedures and establishing their properties.

It is worth noting that Tukey's long running debate with Duncan on this topic was a philosophical one, focusing on what is meant by control of the overall error rate in a situation of multiplicity [(PMC), pages 276–278]. On the other hand his disagreement with Fisher on the use of LSD was more methodological and concerned with the extraction of maximum useful information from the data at hand for a given simultaneous error-rate. He assigned credit to Fisher for recognizing and addressing the problem of simultaneous inference in special situations. Finally, his continuing argument with Scheffé on the use of F-projections had both philosophical and methodological components.

2.3. *Methodology.* Part A of PMC presents an extended discussion of different situations in which the problem of multiplicity arises and illustrates in "cookbook" fashion how these can be treated. Tukey introduces the different error-rates per-X and X-wise (where X stands for experiment or batch) for both significance tests and confidence intervals. He asserts that the practical difference is usually small but that recognizing the difference is important in the development of theory as well as in guiding practice. The latter point is elegantly illustrated in a later discussion on the relative robustness to deviations from normality for per-X and X-wise procedures [(PMC), pages 100–102].

Undoubtedly, Tukey is best known in this field for his introduction of the Wholly Significant Difference procedure (WSD, HSD, T-method) to control the X-wise error rate for a set of multiple comparisons or multiple determinations. In the former case, he demonstrates the connection to the distribution of the Studentized range, which had already been tabulated in part by May (1952) and in the latter case, to the distribution of the Studentized maximum modulus. Using a clever argument, he goes on to show how this control can be easily extended to the case of all contrasts and, ultimately, to all linear combinations, using what he called the Studentized augmented range.

He proceeds to devote substantial effort to developing tables and algorithms, rendering these methods usable by the practitioner. In the process, he derives approximations to the distribution of the Studentized range that extend May's tables beyond 20 determinations and to different error-rates. He also constructs tables of the ratio of the percent points of the WSD to the corresponding percent points of Student's t for different combinations of degrees of freedom and numbers of determinations. While these are of mainly historical interest now, the latter tables do provide some qualitative insight into the effect of taking multiplicity into account. The entire effort is a tour-de-force that illustrates a number of techniques that are now associated with Tukey's work. We return to this point below.

Work in this direction is continued as he employs both analytic and graphical techniques to compare the properties of different methods, such as the WSD with Scheffé's method. Perhaps of greater interest for the contemporary statistician is Tukey's extended treatment of the effect on various methods of departure from normality, principally the introduction of one or more outliers [(PMC), pages 225–240]. This approach combines painstaking analysis and insightful approximations. The results point to a conflict between the forces of efficiency and robustness; Tukey argues that,

> The eventual choice will depend mainly on a judgment of the balance between these two forces.
>
> (PMC), page 241

Still not satisfied with these results, Tukey goes on to examine the effect of inhomogeneity of variance, coming to the conclusion that the procedures that are more robust against nonnormality will also offer some protection against inhomogeneity of variance.

In Part F of PMC, Tukey begins to explore different kinds of simultaneous inference procedures, including the so-called gap-straggler methods, and provides a general framework for generating what he terms "multi-layer significance procedures" that include the Newman–Keuls procedure as well as the procedure of Duncan as special cases. His intent is to clarify why such procedures work well in some circumstances and to encourage the development of methods that have better operating characteristics than those that were available.

2.4. *The tale of Tukey's Conjecture.* Tukey's practical approach to statistical methodology did not explicitly rely on mathematical sophistication. Yet his work has been a continuing source of interesting problems in mathematical statistics, which may be of special interest to the readers of this journal. This is best illustrated by tracing the efforts expended in proving the validity of "Tukey's Conjecture" (first discussed in PMC) concerning the use of the Studentized range in the unbalanced case.

Consider k groups, with n_i observations from a $N(\mu_i, \sigma_i^2)$ population, and let X_i, $i = 1, \dots, k$, be the group means. When all sample sizes and all variances are equal, that is $n_i = n$ and $\sigma_i = \sigma$, a simultaneous set of confidence intervals for all pairwise comparisons of the form $\mu_i - \mu_j$ was given by Tukey using the Studentized range distribution T:

$$X_i - X_j - T_{k,\nu}(\alpha)\sqrt{\frac{1}{n}}S \le \mu_i - \mu_j \le X_i - X_j + T_{k,\nu}(\alpha)\sqrt{\frac{1}{n}}S,$$

where S is the pooled estimator of σ based on ν degrees of freedom.

When the n_i's are not equal, Tukey (1953) suggests a modification to the above set of confidence intervals, replacing n with $(n_i + n_j)/2$. He conjectures that this modification "... is apparently in the conservative direction ...," for all values of k and n_i. Kramer (1956), unaware of Tukey's suggestion, proposed a similar modification (in a somewhat different setting) that, unlike Tukey's work, was published. Hence, the modified procedure for the unbalanced case is referred to as the Tukey–Kramer procedure.

In a doctoral dissertation, under Tukey's supervision, Kurtz (1956) proved the inequality for $k = 3$, and for nearly equal n_i's when $k = 4$. Based on these and other specific configurations he expressed "a strong feeling" for the truth of the conjecture for all k. However, without a mathematical proof, the truth of the conjecture was doubted by others. Miller [(1966), page 87], in his influential book on multiple comparisons, advised against the use of the "inexact" Tukey–Kramer procedure, and suggested instead using the Scheffé procedure.

This overly conservative advice resulted in many efforts to solve the unbalanced pairwise comparisons problem; for example, Spjøtvoll and Stoline (1973), Dunn (1974) and Hochberg (1974, 1975). Nevertheless, the Tukey–Kramer procedure gives shorter intervals, and its validity was further supported by the results of a simulation study by Dunnett (1980). Consequently, in spite of the warnings, it steadily gained popularity among practitioners.

Theoretical interest in the validity of the conjecture remained high and, after 25 years, it began to yield. Using long and complicated proofs, Brown (1979) proved the conjecture for the cases 3, 4 and 5. He then presented it as a difficult dissertation problem to Tony Hayter. Accepting the challenge, Hayter (1984) was able to prove the 31-year-old conjecture for all k in a remarkably short time, even though the argument remained complex. Interestingly, a shorter proof has not been found.

In PMC Tukey also made a more general conjecture that whatever the correlation structure of the estimators of the means, Tukey's procedure always yields conservative confidence intervals. Specifically, consider the case that the vector of means $\mathbf{X}$, which estimates μ has a covariance matrix $\mathbf{V}$, so the variance of $(X_i - X_j)/\sigma^2$ is $d_{ij} = v_{ii} + v_{jj} - 2v_{ij}$. In this setting Tukey's procedure amounts to replacing the term $\sqrt{\frac{1}{n}}$ in the original set of confidence intervals for the balanced case by $\sqrt{d_{ij}/2}$, retaining the same critical value $T_{k,\nu}(\alpha)$.

As in the case of the diagonal matrix, this conjecture has received considerable attention. Hochberg (1974) considered the case when the d_{ij} are all equal, and showed the procedure to be exact. Brown (1984) showed it to hold for $k = 3$. Hayter (1989) extended the proof of the conjecture to those matrices $\mathbf{V}$ for which there exist constants $a_1, a_2, \dots, a_k$ such that $d_{ij} = a_i + a_j$ for all i and j. In spite of such advances the ultimate proof still awaits the courageous investigator.

Even though the conjecture remains to be settled for the comparison of univariate means, a generalized Tukey conjecture has been advanced. It generalizes Tukey's procedure and the relevant conjecture to the pairwise comparisons of multivariate means. Some results and a discussion of the open problems appear in Seo, Mano and Fujikoshi (1994).

2.5. *Observations.* (a) Tukey repeatedly addresses a number of issues of concern to the practitioner. These include choice of error-rate (per-X or X-wise) and choice of procedure. With respect to the latter, early on he makes the point that most of the procedures that had been introduced were admissible so that the choice among them would have to depend on the target of inference as well as on auxiliary optimality criteria.

For example, he argues that if interest centers on pairwise comparisons the WSD is to be preferred to Scheffé's method since it yields substantially shorter confidence intervals. This comparison, elaborated in Tukey (1951), is included in comments on a paper by Scheffé. Implicit in this discussion is the notion of error-rate budgeting and the concomitant advice that it is best to spend one's allowable error-rate wisely by a proper choice of the family of interest or setting priorities among sub-families.

(b) Tukey is willing to make definite, if provisional, recommendations based on the most thorough analysis possible at the time. In fact, most of the last section of PMC, Part F, is devoted to a review and comparison of different procedures (and even families of procedures) with clear guidelines for practice. These include thoughtful warnings on the consequences of the failure of the usual assumptions and prescriptions for how to proceed. Indeed, with few exceptions, it is only in the face of severe departures from normality that Tukey recommends Scheffé's procedure because of the robustness of the F-statistic.

(c) PMC is noteworthy for a number of characteristics we now associate with Tukey, including the heavy use of graphics and clever data analysis. It also addresses a number of issues that occupied Tukey's attention throughout his career, such as the robustness of statistical procedures to failures of the usual assumptions. Tukey's attack on the problem anticipates his well-known paper on sampling from contaminated distributions [Tukey (1960)] as well as the use of sensitivity curves in the Princeton Robustness Study [Andrews et al. (1972)].

(d) In PMC, Tukey expresses strong feelings on the role of the statistician and statistical reasoning in scientific work. Professional statisticians, he believes, have much to learn from the methods of good scientists and also bear an obligation to offer alternatives (or entirely new approaches) that meet real needs and are practical as well. Another obligation of the statistician, as a methodological generalist, is to develop insights of value to the scientific enterprise.

(e) The role of mathematics in statistics is also addressed [(PMC), pages 183–184]. Tukey argues that while it is clearly useful in deriving proofs and insight, mathematics cannot do it all. To paraphrase Tukey, statistics is a part of science and not a branch of mathematics. The discussion of how one is to choose among admissible procedures (referred to above) is an example of where the statistician must draw on more than mathematics to make an appropriate decision.

3. Tukey's later contributions to multiple comparisons.

3.1. *A chronology.* From the early 1950s to the early 1990s Tukey published only two papers on multiple comparisons. Nevertheless, throughout this period Tukey remained convinced that the issue of multiplicity was a fundamental problem of practical importance. This was reflected, for example, in his including topics related to multiple comparisons as a part of his famous Statistics 411 class on Data Analysis at Princeton University, as well as in the Advances in Data Analysis course which he taught intermittently over 20 years to scientists and engineers.

His article in *Science* [Tukey (1977a)] was titled "Some thoughts on clinical trials, especially problems of multiplicity." Highlighted in that paper are the problems of (i) multiple endpoints, (ii) the selection of subsets of patients for which the new treatment seems to offer improvement, and (iii) the question of multiple looks at intermediate points during the course of the experiment ("Not-Very-Sequential designs"). The paper does not introduce any major new idea or methodology, but it emphasizes the importance of a well-designed clinical trial, with a clear protocol specifying in advance the decisions to be taken.

Further evidence of the importance that Tukey attributed to the problem of multiplicity, comes in his discussion of the Bayesian framework for clinical trials: "I have yet to see a Bayesian account in which there is an explicit recognition that the numbers we are looking at are the most favorable out of k. Until I do, I doubt that I will accept a Bayesian approach to questions of this sort as satisfactory" [Tukey (1977a)]. Recent Bayesian efforts in this direction are reviewed by Berry and Hochberg (1999).

While multiplicity was considered by Tukey to be of great practical importance, at this stage he appeared to feel that the philosophy and methodologies were quite mature, and little remained to be done in this area—a point he explicitly made in his 1977 talk at the Annual Meeting of the American Statistical Association. However, as one can see from his Statistics 411 class notes, he never stopped trying new ideas. For instance, he elaborated a pedagogical example from the PMC, pages 74–77, into "The Higher Criticism." On the other hand, handouts from that period discussing the higher criticism [Tukey (1977b) and Tukey, Bloomfield, Braun and McNeill (1978)] never became full fledged studies, suggesting that he did not believe this was a fruitful avenue to pursue. Interestingly, this approach has been explored by Donoho and Jin (personal communication), yielding some asymptotic results.

In 1983 Tukey made another excursion into the field of multiple comparisons, triggered by an invitation to write for the Lord festschrift volume. In this work, Braun and Tukey (1983) explored the implications of recent developments in the area of stepwise multiple testing, by Marcus, Peritz and Gabriel (1976), Begun and Gabriel (1981) and Ramsey (1981), while building upon older ideas in the problem of pairwise comparisons. The goal was typical for Tukey: to develop stepwise methods simple enough for use in practice.

By 1989 he was again deeply involved with multiple comparisons. His renewed interest, probably sparked by the invitation to deliver the Rupert G. Miller Memorial Lecture, resulted in a presentation which bore the title “The philosophy of multiple comparisons” [Tukey (1991)]. It is also possible that this interest grew from his involvement at the time in the statistical analyses of two substantial data sets, characterized by large scale multiplicity problems. The first was the analysis of data from NAEP, involving all pairwise comparisons among the states. The second was the analysis of breeding experiments in which a multitude of properties are compared among many strains, planted at many different sites.

In chronological order, following “The philosophy of multiple comparisons” came “Graphic comparisons of several linked aspects” [Tukey (1993a)] which is devoted mainly to the display of the results of simultaneous analysis, and “Where should multiple comparisons go next?” [Tukey (1993b)]. Tukey had become interested in the new criterion of the False Discovery Rate [Benjamini and Hochberg (1995)], and this was reflected in the studies that followed. See also Tukey’s remarks in Tukey (1995) and in Brillinger, Fernholz and Morgenthaler (1997). The analysis of educational data resulted in a 1994 technical report by Williams, Jones and Tukey, finally published only in 1999, and in “Controlling the proportion of false discoveries for multiple comparisons—future directions” [Tukey (1995)]. The first part of the analysis of the multiplicity problems in breeding experiments was completed in 1995 [appearing in print in Basford and Tukey (1997)] and culminated in the publication by Basford and Tukey (1998).

Through his last year he continued to work and write on these and related issues: “Improved multiple comparison procedures for controlling the false discovery rate” with Charles Lewis [Lewis and Tukey (2001)], “A sensible formulation of the significance test” with Lyle Jones [Jones and Tukey (2000)] and a somewhat unconventional encyclopedia entry on multiplicity [Jones, Lewis and Tukey (2001)]. In the following sections we discuss the principal themes of this body of work with little attention to chronological order.

3.2. *The purpose of inference.* Over the last two decades the practice of testing statistical hypotheses has become increasingly controversial. Hypothesis testing is often contrasted with different, recommended practices of statistics. By favoring confidence intervals (or exploratory data analysis, graphical methods, Bayesian statistics, etc.), practitioners implicitly reject the use of significance tests.

Tukey’s approach was different. Objecting to the point hypotheses treated in the Neyman–Pearson formulation, he went to great lengths to warn that pure hypothesis testing is unrealistic: the “null hypothesis” never holds, at least to some decimal point. Therefore, an “error” does not occur as a result of erroneously rejecting a never true null hypothesis. On this issue he sounded his infamous observation: “Statisticians classically asked the wrong question—and were willing to answer with a lie." [Note: All quotations in this section are taken from Tukey (1991), page 100.] At the same time, he valued significance testing as an important practical tool. How could the two points of view be reconciled? By viewing hypothesis testing through the lens of the three-decisions-rule: "What we should be answering first is 'Can we tell the direction in which the effect of A differs from the effect of B?' In other words can we be confident about the direction from A to B? Is it 'up,' 'down' or 'uncertain'?"

It might appear that this is just a matter of rewording. It was not for Tukey: "Unless we learn to keep what we say, what we think and what we do all matching one another, and matching a reasonable picture of the world, we will never find our way safely through the thickets of multiple comparisons." He certainly adhered to the consequences of this approach. In his later writings he repeatedly referred to 5%, as the 2.5% directional error, and therefore was satisfied with an overall level of 10%. He also suggested the "perinull situation," which models negligible effects by a low variance normal distribution centered at 0, and used it in simulation studies.

Still, confidence intervals were viewed as the more important inferential tools, primarily because they excluded a large set of values from the set of possible values for the parameters. But even here his ideas about the role of confidence intervals were somewhat unconventional. He treated the roles of the two ends of the confidence interval differently. If $A - B$ was definitely (significantly) positive, then the "larger part of the follow-up question is what is the minimum size of $A - B$. The smaller part, usually, answers: what is the maximum size of $A - B$." It may be of interest to relate these ideas to the growing theoretical interest in nonequivariant confidence intervals, which can be tailored to the specific target of inference [Hayter and Hsu (1994), Brown, Casella and Hwang (1995) and Benjamini, Hochberg and Stark (1998)].

Why should the goal of establishing confidence intervals ever be compromised by settling for confident directions? Two answers are given [in Tukey (1991) and in Jones and Tukey (2000)]: The first argument is that sometimes the scale on which A and B are compared is not quite fully developed and does not yet have an interpretation of its own. Thus a quality-of-life questionnaire, summarized on a scale of 1 to 34 may be enough to establish which of two treatments is better, but has no importance or relevance of its own. The second argument is again practical, depending on the fact that by restricting the inference to confident directions more sensitive analyses are feasible.

3.3. *The hierarchy of error-rates.* Almost every time Tukey wrote about multiple comparisons he used the opportunity to emphasize the importance of controlling either the per family (per batch or per experiment) or the familywise (batchwise or experimentwise) error-rate. Tukey viewed the difference between these two as "a far lesser point," and he was prepared to "neglect this difference whenever neglect simplifies our discussion" [Tukey (1993b), pages 189–190], in sharp contrast to the per comparison (comparison-wise) approach, which amounts to ignoring the multiplicity problem altogether.

He believed, and often expressed the opinion, that these concepts were not easy to understand even by statisticians (at this point the reader is encouraged to conduct a self-examination), let alone by other scientists who practice statistics. In each paper Tukey explained the relevant concepts in a somewhat different way, as if hoping that one of these variations might relieve the readers' difficulty.

While remaining the most prominent proponent of simultaneous control, Tukey always showed interest in compromise based on rational analysis. In Jones and Tukey (2000) he explicitly says, "There has been a feeling for decades [e.g., PMC (1953)] that 'individual' or 'unadjusted' is too soft while simultaneous is too severe." It is not apparent from our reading that this belief is clearly expressed in his earlier work, but it is evident in his later work, especially as he attended to larger practical problems.

In this vein, we would argue that Tukey's adoption of the stepwise methods of Welsch (1977) and Ramsey (1981) represented a form of compromise. Since stepwise procedures allow only significance statements (i.e., confident directions only), they represent a relaxed form of control, which is tolerated only because of its increased resolution. Braun and Tukey (1983) presented an approach to achieving the same goal by simpler means. A decade later, Tukey (1993b) suggested linking the goals of simultaneous inference to the magnitude of the ratio of the usual size of a difference to the typical standard error. Strong simultaneous inference (in the form of the Studentized range) is needed when this ratio is both very large and very small. In intermediate cases stepwise testing procedures were recommended.

Some versions of the graphical methods which he devised in later years also attest to his search for a balance between stringent multiplicity control and concern about power. He proposes the display of both individual intervals and simultaneous ones on the same plot. He also encourages support of the usual simultaneous 95% interval with simultaneous 50%, as a way of extracting more "hints" from the data [Tukey (1993a)]. His most recent attempt in this direction was the use of $m^{2/3}$ instead of m in the Bonferroni pairwise adjustment, which involves the divisor $m(m-1)/2$ [Hoaglin, Mosteller and Tukey (1991)].

The false discovery rate (FDR), offered by Benjamini and Hochberg (1995) typifies such an intermediate approach. Tukey learned about the new error-rate as early as 1990, and initially offered slight modifications to its definition, finally settling on the original definition [Jones, Lewis and Tukey (2001)], except for what he regarded as an error. In the original definition a false discovery is the rejection of a true null hypothesis. For Tukey, an error occurs in making a definite statement about the direction of a difference, when the true direction is in fact the opposite. Thus, the FDR could stand for "False Definite Rate" (or False Directional error Rate), as a type I error is not considered possible. Incidently, a type II error is also not considered a real "error"—just a missed opportunity for making a definite statement.

By 1993 Tukey and his collaborators tried the FDR-controlling procedure on both the NAEP data and the breeding data, and appreciated its operating characteristics. He adopted the approach and in the years that followed contributed his own ideas to FDR methodology. His preference was based on the following rationale [Jones, Lewis and Tukey (2001)]:

> In a situation where only very few of the results are definite, the "false discovery" approach performs much like the simultaneous approach. Since these situations include those where "definite" is purely accidental, as in extreme situations of data mining, it is important to have a severe procedure. In a situation in which most comparisons earn a definite result, the FDR algorithm behaves rather like the individual (i.e., unadjusted) procedure. Again if many comparisons deserve a definite result, it is reasonable that looking for the most extreme is not going to lead to excessively many "errors."

3.4. *The hierarchy of families in ANOVA.* In his early work, Tukey emphasized four families or sets of inferences of importance, all of which have already been mentioned: the set of determinations, the set of pairwise comparisons, the set of all contrasts and the set of all linear combinations. To a lesser extent he was interested in the set of comparisons with a single control, although he did make a late contribution to the area [Almond, Lewis, Tukey and Yan (2000)]. His later work in data analysis of variance, culminating in the book by Hoaglin, Mosteller and Tukey (1991), awakened interest in other natural families of practical importance in ANOVA. With respect to the two factor case, one might be interested in the two sets of pairwise comparisons among main effects, the interactions, the pairwise comparison of interactions, the conditional comparisons within a row or a column, and so on.

It is true that in a two-way ANOVA setting with replications, viewing the problem as that of comparing IJ samples each with mean μ_{ij}, these questions can be answered by applying the Studentized range to contrasts and linear combinations. Tukey, however, was interested in utilizing the structure of more limited families to get higher sensitivity. Thus he suggested looking separately at the family of bicomparisons, which are comparisons of the form $(y_{ai} - y_{bi}) - (y_{aj} - y_{bj})$ for all $a < b$ and $i < j$, and went on to study the theoretical distribution of the maximum of the absolute size of the such Studentized deviations. We have no direct reference to such work, but Hoaglin, Mosteller and Tukey (1991) give a working approximation using the easily available Studentized range distribution.

Even in the pairwise comparisons of main effects, he attempted to gain more sensitivity by studying families defined by deviations from a pooled 25% trimmed mean cluster. In view of his interest in a richer collection of families of inference, he found the FDR approach a relief in that it offered "less dependency on the exact definition of the family" [Basford and Tukey (1998)].

3.5. *Pooling.* On various occasions Tukey advocated the method of "Leaving out or pooling noisy estimates." In a large breeding experiment, clustering was used to join such effects into many fewer groups. This was not meant to imply that the means within a group are equal, but rather that we do not know in what direction the difference lies, so we might as well treat them as a batch. For Tukey "the importance of grouping as an aid to clarity" [Tukey (1993b), page 195] was paramount.

The same idea took the form of "sweeping" lines in complex ANOVA tables [e.g., Hoaglin, Mosteller and Tukey (1991), Chapter 11]. Tukey considered this concept as fundamental, and expressed it as "The $F > 2$ principle": Collapse lines in ANOVA once they are below this threshold. In view of his later writing, we believe that "$F > 2$" should be treated more as a method, and the "Leaving out or pooling noisy estimates" as the fundamental principle. This is very much in keeping with recent developments in decision theory, that attest to the advantages of thresholding—the setting to zero of noisily estimated coefficients. Interestingly, Donoho and Johnstone's (1994) global threshold has a simultaneous inference interpretation as a Bonferroni-adjusted testing procedure.

3.6. *Pairwise comparisons using FDR.* Tukey was the first to note that the FDR approach and the linear step-up procedure in Benjamini and Hochberg (1995) (denoted by him as the B-H procedure or the BSD) could be very useful in the pairwise comparisons situation, where the multiplicity problem becomes large even when a moderate number of groups is compared. Since the original proof of control of FDR assumed independence, Williams, Jones and Tukey (1994, 1999) conducted an extensive simulation study to verify that the B-H procedure controls the FDR in the pairwise setting. Tukey was clearly satisfied that the results provided sufficient assurance that the FDR is conservatively controlled by the B-H procedure and the authors sum up their discussion in the following way: "Each of the three authors believes that the B-H procedure is the best available choice."

Other studies [Benjamini, Hochberg and Kling (1993) and Keselman, Cribbie and Holland (1999)] give additional support for the conjecture. Theoretical progress in proving the probability inequality assuring FDR control in the normal pairwise balanced setting has been limited [Yekutieli (2001)]. It is possible that this problem will remain open for many years (as was the case with Tukey's conjecture), yielding only after a number of small gains are achieved. A theoretical answer to the current problem is especially desirable, as it may improve the procedure. Once this conjecture is settled, the next natural question concerns FDR control in the unbalanced case when using the Tukey–Kramer modification.

3.7. *Graphical displays of simultaneous confidence intervals.* It is no exaggeration to say that throughout his career Tukey was preoccupied with the issue of finding the appropriate graphical tools to compare the means in a one-way analysis. He returned repeatedly to the issue, each visit resulting in a new variation—and sometimes an entirely different approach. In PMC, pages 96–97, he introduced the graphical display of simultaneous confidence intervals for determinations, as well as "allowances" for the comparisons. For comparisons the basic idea is that if there exists a distance beyond which the two means are considered separated, then an effective graphical display involves drawing an allowance equal to plus or minus half that distance around the mean, and noting whether the allowances of the pair of means being compared overlap. (Note that the efficacy of this approach depends on equal distances for all pairs, otherwise the simple approach leads to very interesting theoretical problems.) A number of variations are described in Almond, Lewis, Tukey and Yan (2000).

But Tukey never viewed his latest suggestion as entirely satisfactory. The evolving terminology testifies to his perennial quest: Allowances were followed by "notches" [McGill, Tukey and Larsen (1978)] which were in turn called interferences, followed by the use of "gauges" and "overlaps" [Basford and Tukey (1997)]. Even after he devoted a long manuscript to the display of simultaneous determinations and comparisons, studying in detail dozens of variations [Tukey (1993a)], upon his next visit to the subject we find yet another display [Williams, Jones and Tukey (1994, 1999)]. This display is based on the 45° tilting with which he experimented semigraphically in Tukey (1991). [It is interesting to note here that Hsu (1996) has created some powerful graphical displays by adding graphical confidence intervals for comparisons and introducing a high level of graphical interaction.] Finally, in Basford and Tukey (1998) the new "staircase display" makes it to the cover of the book. The groups' means and "overlaps" are displayed in such a way, that for a specific strain of interest, we are confident that any other strain displayed in the lower staircase and to the left has lower yield, and any strain displayed above and to the right has higher yield.

How can this never-ending struggle be explained and what lessons does it hold? Most simply, it may be that trying to present information about the means of k groups, as well as about their $k(k-1)/2$ comparisons in the same display, is too ambitious a goal. This may well be the case, as in later work Tukey separates the two types of displays. A deeper explanation may be derived from his introductory comments in Tukey (1993a): "Graphs should report the results of careful data analysis—rather than be an attempt to replace it." In fact, his thinking about what could or should be represented as a result of the analysis kept evolving over the years: Confidence intervals versus confidence directions; error-rates at 95% or also at 50% and 5%; simultaneous alone or simultaneous and individual or the FDR. Tukey once said regarding multiple comparisons that "there may be a man for all seasons, but there isn't a procedure for all purposes" [Tukey (1991)]. His work on graphical procedures for multiple comparisons is testimony to that dictum.

3.8. *The scope of the field of multiple comparisons.* Tukey opened his last review [Jones, Lewis and Tukey (2001)] by stating that:

> Questions arising from problems of multiplicity raise a diversity of issues, issues that tend to be important, difficult, and often unresolved.

Later, the scope of the field is described as follows:

> Issues of multiplicity appear in a wide variety of kinds of situations, including:
>
> - Comparisons of all kinds of situations ("multiple comparisons").
> - Comparisons between each situation and the standard ("multiple determinations").
> - Selection of one or more candidates from several or many.
> - Selection of how variables are to be expressed for analysis.
> - Selection of which of the available variables are to participate in our analysis (subset selection).
> - Reduction of a collection of candidates in an attempt to reduce their number without losing the best.

Even though he chose in that review to address only one case, the familiar issue of pairwise comparisons, he saw not only a broader agenda for the field, but also tried to emphasize its practical importance and conceptual difficulties. In non-Tukey terminology we find here the problems of multiple parameter testing and confidence estimation, the pairwise comparison problem, many-to-one comparisons, ranking and selection, model selection, variable selection, high throughput screening, etc. He thus leaves us with more than enough topics to grapple with for the decades ahead—especially in view of his usual advice that there need not be a single best answer.

4. Conclusions.

As in so many other areas of statistics, it is not difficult to argue that Tukey played a seminal role in the development of simultaneous inference. PMC invested the field with both academic respectability and a focus on practical issues. Through vigorous discussion with his statistical colleagues, he helped to clarify the terms of the debate and identified key issues. Over time, many of Tukey's views on both philosophy and methodology have carried the day among researchers on multiplicity. Even though the debate is not over in the wider community, the importance of the simultaneous control of error and the role of error-rate budgeting is a recognized practice. His analysis of multilayer procedures led directly to theoretical investigations of step-wise approaches, arguably the most productive area of research over the 35 years following the publication of PMC. See Hochberg and Tamhane (1987). Many researchers built directly on the insights and approaches pioneered in PMC. Certainly, Tukey's concerns with robustness and the use of graphical methods (which transcend multiplicity) have become major areas of investigation.

It is difficult to predict how Tukey's work over the last decade will influence the field in years to come. Unquestionably, he was prescient in beginning to address real problems involving large data sets where multiplicity issues cannot be ignored. While he attended to the analysis of education data, statisticians and others are now engaged in a wide variety of problems, ranging from data mining in market research to the analysis of genomics studies. Perhaps more important than the particular techniques that he proposed is the standard he set for all scientists: An ongoing willingness to reexamine assumptions, to experiment with new approaches (one's own or that of others) and to seek practical methods informed by theoretical insights.

Indeed, the quest for practical solutions to real problems has long been the lodestar of Tukey's work right through his last efforts. A number of subject area reviews of multiplicity procedures have appeared in recent years [e.g., Ottenbacher (1998), Keselman, Cribbie and Holland (1999), Wilkinson (1999) and Curran-Everett (2001)]. Despite the fact that a good deal of Tukey's reputation among experimentalists rests on his contributions to multiplicity, these reviews indicate that there is great heterogeneity in the degree of penetration of ideas about multiplicity into general practice. Even in areas where such procedures are used as a matter of course, a few relatively simple methods account for most of the applications. Thus, while considerable progress has been made, there is still a long road to follow.

Tukey was not unaware of this situation and he adopted two complementary approaches in response. First, he engaged in continuing efforts to reach working scientists and engineers with instructive discussions of the problems of multiplicity and ways of dealing with them. Second, he focused on how to develop a comprehensive approach to the analysis of large, complex data sets in which multiplicity plays an important but not necessarily central role. In this vein, the extended treatment by Basford and Tukey (1998) of a plant breeding trial should prove to be a milestone in the thoughtful integration of a variety of statistical techniques in the analysis of data.

Given Tukey's track record over nearly 60 years of statistical work, we would argue that everyone can benefit from the rich body of work he left behind. Mathematical statisticians, in particular, should consider both the philosophical issues and practical matters that preoccupied Tukey. They encompass enough open problems, conjectures, hints and ambiguities to keep many of us occupied for years to come. We venture to say that Tukey would be most pleased if his legacy were to inspire the next generation of investigators and lead to more powerful and useful techniques that can be applied in the quest for scientific understanding.

Acknowledgments. The authors would like to thank Charles Lewis for a detailed review and Yossi Hochberg for some historical information.

REFERENCES

ALMOND, R. G., LEWIS, C., TUKEY, J. W. and YAN, D. (2000). Displays for comparing a given state to many others. *Amer. Statist.* **54** 89–93.

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location*: *Survey and Advances.* Princeton Univ. Press.

BASFORD, K. E. and TUKEY, J. W. (1997). Graphical profiles as an aid to understanding plant breeding experiments. *J. Statist. Plann. Inference* **57** 93–107.

BASFORD, K. E. and TUKEY, J. W. (1998). *Graphical Analysis of Multiresponse Data*. Chapman and Hall, London.

BEGUN, J. and GABRIEL, K. R. (1981). Closure of the Newman–Keuls multiple comparisons procedure. *J. Amer. Statist. Assoc.* **76** 241–245.

BENJAMINI, Y. and 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.

BENJAMINI, Y., HOCHBERG, Y. and KLING, Y. (1993). False discovery rate control in pairwise comparisons. Working Paper 93-2, Dept. Statistics and O.R., Tel Aviv Univ.

BENJAMINI, Y., HOCHBERG, Y. and STARK, P. B. (1998). Confidence intervals with more power to determine the sign: Two ends constrain the means. *J. Amer. Statist. Assoc.* **93** 309–317.

BERRY, D. A. and HOCHBERG, Y. (1999). Bayesian perspectives on multiple comparisons. *J. Statist. Plann. Inference* **82** 215–227.

BRAUN, H. I. and TUKEY, J. W. (1983). Multiple comparisons through orderly partitions: The maximum subrange procedure. In *Principals of Modern Psychological Measurement*: *A Festschrift for Frederic M. Lord* (H. Wainer and S. Messick, eds.) 55–65. Erlbaum, Hillsdale, NJ.

BRILLINGER, D. R., FERNHOLZ, L. T. and MORGENTHALER, S., eds. (1997). *The Practice of Data Analysis.* Princeton Univ. Press.

BROWN, L. D. (1979). A proof that the Tukey-Kramer multiple comparison procedure for differences between treatment means is level-α for 3, 4, or 5 treatments. Technical report, Dept. Mathematics, Cornell Univ.

BROWN, L. D. (1984). A note on the Tukey–Kramer procedure for pairwise comparisons of correlated means. In *Design of Experiments*: *Ranking and Selection* (*Essays in Honor of Robert E. Beckhofer*) (T. J. Santner and A. C. Tamhane, eds.) 1–6. Dekker, New York.

BROWN, L. D., CASELLA, G. and HWANG, J. T. G. (1995). Optimal confidence sets, bioequivalence, and the limaçon of Pascal. *J. Amer. Statist. Assoc.* **90** 880–889.

CURRAN-EVERETT, D. (2001). Multiple comparisons: Philosophies and illustrations. *Amer. J. Physiology*: *Regulatory Integrative and Comparative Physiology* **279** R1–R8.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

DUNN, O. J. (1974). On multiple tests and confidence intervals. *Comm. Statist.* **3** 101–103.

DUNNETT, C. W. (1980). Pairwise multiple comparisons in the homogeneous variance, unequal sample size case. *J. Amer. Statist. Assoc.* **75** 789–795.

HAYTER, A. J. (1984). A proof of the conjecture that the Tukey–Kramer multiple comparisons procedure is conservative. *Ann. Statist.* **12** 61–75.

HAYTER, A. J. (1989). Pairwise comparisons of generally correlated means. *J. Amer. Statist. Assoc.* **84** 208–213.

HAYTER, A. and HSU, J. (1994). On the relationship between stepwise decision procedures and confidence sets. *J. Amer. Statist. Assoc.* **89** 128–136.

HOAGLIN, D. C., MOSTELLER, F. and TUKEY, J. W., eds. (1991). *Fundamentals of Exploratory Analysis of Variance*. Wiley, New York.

HOCHBERG, Y. (1974). The distribution of the range in general balanced models. *Amer. Statist.* **28** 137–138.

HOCHBERG, Y. (1975). An extension of the t-method to general unbalanced models of fixed effects. *J. Roy. Statist. Soc. Ser. B* **37** 426–433.

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

HSU, J. C. (1996). *Multiple Comparisons*: *Theory and Methods.* Chapman and Hall, London.

JONES, L. V., LEWIS, C. and TUKEY, J. W. (2001). Hypothesis tests, multiplicity of. In *International Encyclopedia of the Social and Behavioral Sciences* (N. J. Smelser and P. B. Baltes, eds.) 7127–7133. Elsevier, London.

JONES, L. V. and TUKEY, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods* **5** 411–414.

KESELMAN, H. J., CRIBBIE, R. and HOLLAND, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise type I error control. *Psychological Methods* **4** 58–69.

KRAMER, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* **12** 307–310.

KURTZ, T. E. (1956). An extension of a multiple comparisons procedure. Ph.D. dissertation, Princeton Univ.

LEWIS, C. and TUKEY, J. W. (2001). Improved multiple comparison procedures for controlling the false discovery rate. Unpublished manuscript.

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.

MAY, J. M. (1952). Extended and corrected tables of the upper percentage points of the "Studentized" range. *Biometrika* **39** 192–193.

MCGILL, R., TUKEY, J. W. and LARSEN, W. O. (1978). Variations on box plots. *Amer. Statist.* **32** 12–16.

MILLER, R. G. (1966). *Simultaneous Statistical Inference*. McGraw-Hill, New York.

OTTENBACHER, K. J. (1998). Quantitative evaluation of multiplicity in epidimiology and public health research. *Amer. J. Epidimiology* **147** 615–619.

RAMSEY, P. H. (1981). Power of univariate pairwise multiple comparison procedures. *Psychological Bulletin* **90** 352–366.

SEO, T., MANO, S. and FUJIKOSHI, Y. (1994). A generalized Tukey conjecture for multiple comparisons among mean vectors. *J. Amer. Statist. Assoc.* **89** 676–679.

SHAFFER, J. P. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology* **46** 561–584.

SPJØTVOLL, E. and STOLINE, M. R. (1973). An extension of the T-method of multiple comparison to include the cases with unequal sample sizes. *J. Amer. Statist. Assoc.* **68** 975–978.

TUKEY, J. W. (1951). Reminder sheets for "Discussion of paper on multiple comparisons by Henry Scheffé." In *The Collected Works of John W. Tukey VIII. Multiple Comparisons*: *1948–1983* 469–475. Chapman and Hall, New York.

TUKEY, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript. In *The Collected Works of John W. Tukey VIII. Multiple Comparisons*: *1948–1983* 1–300. Chapman and Hall, New York.

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*: *Essays in Honor of Harold Hotelling* (I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann, eds.) 448–485. Stanford Univ. Press.

TUKEY, J. W. (1977a). Some thoughts on clinical trials, especially problems of multiplicity. *Science* **198** 679–684.

TUKEY, J. W. (1977b). Higher criticism for individual significances in several tables or parts of tables. Internal working paper 89-9, Princeton Univ.

TUKEY, J. W. (1991). The philosophy of multiple comparisons. *Statist. Sci.* **6** 100–116.

TUKEY, J. W. (1993a). Graphic comparisons of several linked aspects: Alternatives and suggested principles (with discussion). *J. Comput. Graph. Statist.* **2** 1–49.

TUKEY, J. W. (1993b). Where should multiple comparisons go next? In *Multiple Comparisons, Selection, and Applications in Biometry* (F. M. Hoppe, ed.) 187–207. Dekker, New York.

TUKEY, J. W. (1994). *The Collected Works of John W. Tukey VIII. Multiple Comparisons*: *1948–1983*. Chapman and Hall, New York.

TUKEY, J. W. (1995). Controlling the proportion of false discoveries for multiple comparison—Future directions. In *Perspectives on Statistics for Educational Research*: *Proceedings of a Workshop* (V. S. Williams, L. V. Jones and I. Olkin, eds.). Technical Report 35, National Institute of Statistical Sciences, Research Triangle Park, NC.

TUKEY, J. W., BLOOMFIELD, P., BRAUN, H. I. and MCNEILL, D. R. (1978). Advances in data analysis. Unpublished manuscript.

TUKEY, J. W., CIMINERA, J. L. and HEYSE, J. F. (1985). Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics* **41** 295–301.

WELSCH, R. E. (1977). Stepwise multiple comparison procedures. *J. Amer. Statist. Assoc.* **72** 566–575.

WILKINSON, L. (1999). Statistical methods in psychology journals—guidelines and explanations. *American Psychology* **54** 594–604.

WILLIAMS, V. S. L., JONES, L. V. and TUKEY, J. W. (1994). Controlling error in multiple comparisons, with special attention to the National Assessment of Educational Progress. Technical Report 33, National Institute of Statistical Sciences, Research Triangle Park, NC.

WILLIAMS, V. S. L., JONES, L. V. and TUKEY J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics* **24** 42–69.

YEKUTIELI, D. (2001). Controlling the false discovery rate under dependency. Ph.D. dissertation, Dept. Statistics, Tel Aviv Univ. (in Hebrew).

DEPARTMENT OF STATISTICS
TEL AVIV UNIVERSITY
RAMAT-AVIV
ISRAEL

EDUCATIONAL TESTING SERVICE
ROSEDALE ROAD
PRINCETON, NEW JERSEY 08541
E-MAIL: hbraun@ets.org