Colloquy: Should Familywise Alpha Be Adjusted?
Against Familywise Alpha Adjustment

DANIEL J. O'KEEFE
University of Illinois at Urbana-Champaign

Type I error is a risk undertaken whenever significance tests are conducted, and the chances of committing a Type I error increase as the number of significance tests increases. But adjusting the alpha level because of the number of tests conducted in a given study has no principled basis, commits one to absurd beliefs and practices, and reduces statistical power. The practice of requiring or employing such adjustments should be abandoned.

When a large number of statistical significance tests are performed in a given study, there is a common requirement that the alpha level (the Type I error rate) be adjusted to reflect the number of tests being conducted. This article argues that requiring or employing such alpha adjustments is unjustified and undesirable, and hence that such adjustments should not be undertaken.

BACKGROUND

Whenever a researcher undertakes a large number of statistical significance tests in a given study, critical readers will commonly be alert to the possibility that any obtained significant effects might reflect Type I error (i.e., incorrectly rejecting a true null hypothesis). With an increase in the number of significance tests conducted (over data where the null hypothesis is true), a correspondingly larger number of significant results will be obtained simply because of Type I error. Hence, a variety of statistical procedures have been developed that permit an investigator to control for this problem. Broadly speaking, these procedures involve reducing

---

Daniel J. O'Keefe (Ph.D., University of Illinois, 1976) is a professor in the Department of Speech Communication at the University of Illinois at Urbana-Champaign. He thanks Dale Brashers, John Caughlin, Nosh Contractor, Sally Jackson, Scott Jacobs, Bruce Lambert, Robin Nabi, Barbara O'Keefe, and faculty and students in the Department of Communication at the University of Arizona. Correspondence concerning this article should be addressed to Daniel J. O'Keefe, Department of Speech Communication, 702 S. Wright St., 244 Lincoln Hall, University of Illinois at Urbana-Champaign, Urbana, IL 61801-3694, USA; email: dokeefe@uiuc.edu.

Human Communication Research, Vol. 29, No. 3, July 2003 431–447

the alpha level (below its customary .05 value) so as to produce the desired overall level of protection against Type I error.

One common alpha-adjustment procedure is the Bonferroni method, in which the alpha level to be used is obtained by dividing the desired overall alpha level (say, .05) by the number of tests to be performed. For example, if a researcher is performing eight tests and wants an overall alpha of .05, alpha is set to .00625 for each of the eight tests. Thus the "familywise" alpha level (the error rate for the entire collection or family of tests) is .05; the "per comparison" alpha level (the error rate for each test) is .00625.

Bonferroni methods provide just one example of alpha-adjustment procedures. Quite a number of different alpha-adjustment procedures (variously labeled) are available, including Tukey (honestly significant difference), Duncan (multiple range), Newman-Keuls (Student-Newman-Keuls), Dunnett, and Scheffe[1]. The varieties and details, however, can be put aside here, because the present argument is meant to apply quite generally to such procedures.

As already intimated, the rationale for the application of these procedures invokes reasoning along the following lines: As the number of tests increases, the chance of making a Type I error increases. With the usual .05 alpha level, one expects 5% of the tests to be significant by chance alone when the null hypothesis is true. So, for instance, a researcher who conducts 100 tests and finds five significant effects should (by this reasoning) be worried about having "capitalized on chance," and certainly should not be permitted to interpret the significant effects straightforwardly. To control for this inflation in Type I error, alpha-adjustment procedures are used.

This reasoning is commonplace. For example, Rosenthal and Rosnow (1984) note that "generally, the more tests of significance computed on data for which the null hypothesis is true, the more significant results will be obtained, i.e., the more type I errors will be made" (p. 254). Therefore, in circumstances in which "investigators are worried lest they are capitalizing on chance, a simple and quite conservative procedure can be employed to adjust the interpretation of the *p* values obtained . . . [namely] the Bonferroni procedure" (p. 255).

Indeed, discussions of multiple-comparison procedures almost inevitably invoke the specter of Type I error rate inflation. For example: "The major problem resulting from the performance of a series of analytical comparisons on a set of data is the unpleasant fact that the more comparisons we conduct, the more type I errors we will make when the null hypothesis is true" (Keppel, 1991, p. 164). Or: "when each of a set of comparisons is to be tested, then we should also be concerned over the probability of making *at least one* Type I error in the entire set or family of comparisons. This probability of making one or more Type I errors in the

set of comparisons is known as the familywise error rate, or FW error rate" (Hays, 1988, p. 410). Or: "Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true" (Moore & McCabe, 1999, p. 481).

Fear of such Type I error rate inflation thus motivates the use of alpha-adjustment procedures. For example: "control for excessive familywise error" can be accomplished by "setting a maximum familywise error rate using one of the Bonferroni principles" (Hays, 1988, p. 411). Or:

> if we . . . want to take into account the fact that as the number of planned contrasts increases the probability of a type I error also increases, we might want to adjust our alpha (significance) level accordingly. A simple and generally conservative way to make this adjustment is based on the Bonferroni approach. (Rosenthal & Rosnow, 1985, p. 45)

Although the use of these alpha-adjustment procedures is most commonly found in treatments of experimental research data, one can see similar procedures urged (with a similar underlying principle) in other research domains. For example, Matt and Cook (1994, p. 508) have argued that "meta-analysts may capitalize on chance . . . by conducting a large number of statistical tests without adequately controlling for Type I error."

In brief, then, the general principle underlying these alpha-adjustment procedures is that as the number of significance tests increases, the alpha level must be adjusted correspondingly to avoid capitalizing on chance.

THE CASE AGAINST ALPHA ADJUSTMENT

The case against using or requiring such alpha-adjustment procedures can be summarized in three observations: Such procedures reduce statistical power, the principle justifying such procedures is not consistently applied, and consistent application of that principle would commit one to bizarre beliefs and undesirable research practices.

Reduced Statistical Power

Alpha-adjustment procedures reduce statistical power. Obviously, *ceteris paribus*, setting a lower alpha level will mean reducing the chances of detecting a genuine (nonzero) population effect. Such reduced power is not itself a compelling reason to abandon alpha-adjustment procedures and is not offered as such. Rather, the point of this initial observation is to make it plain that these procedures do incur some costs (namely, overlooked effects). This, in turn, implies that any imposition of these procedures requires some justification.

Inconsistent Application

The principle commonly used to justify alpha-adjustment procedures is not consistently invoked. The usual application of these procedures is found in experimental designs with multiple independent variables and hence with multiple possible comparisons among cells. For example, a 3 x 2 design contains six cells and so provides 15 possible pairwise comparisons of cell means. If a researcher conducts all 15 comparisons, critical readers will commonly insist on the use of an alpha-adjustment procedure that takes into account the number of tests being performed.

But now consider the significance tests associated with multiple-regression analyses. Imagine that an investigator begins by entering a block of two predictor variables, each tested for significance. On the next step, three more predictors are entered, so that at this second step five predictors are tested, giving a total of seven significance tests performed thus far. On the next step, three more predictors are added, so that at this step eight significance tests are performed, making a total of 15 significance tests conducted across the three steps. Yet investigators almost never, and are almost never asked to, adjust alpha to take into account the number of tests performed in such an analysis.

Similarly, consider the case of an investigator who assesses the significance of the intercorrelations among six variables. There are 15 such correlations, but investigators typically do not, and are not typically asked to, adjust alpha so as to take into account the number of tests performed.

Or consider the significance tests ordinarily conducted in an initial analysis of variance. In a factorial design with four independent variables, there are 15 effects—four main effects and 11 interactions—and so 15 significance tests. Yet most researchers do not adjust, nor are they commonly asked to adjust, the alpha level in such an ANOVA so as to take into account the number of tests being performed.

What these examples suggest is that at a minimum the policy of correcting alpha for the number of tests conducted is not currently applied consistently. If the rationale for adjusting alpha is the avoidance of Type I error occasioned by performing a large number of statistical tests, then whenever a large number of statistical tests is performed, alpha adjustment should be required. That it is not gives rise to doubts about whether the basis for imposing these alpha-adjustment procedures is genuinely "principled."

Consequences of Consistent Application

Consistent application of the principle justifying alpha-adjustment procedures ("as the number of significance tests increases, the alpha level must be adjusted correspondingly") would have absurd and undesirable

consequences. As an initial illustration of the absurdities attendant to consistent application of the general principle, consider the case in which a researcher reanalyzes another researcher's data set. Suppose, for example, that Professor P and I have differing views on the relationship of variable X and variable Y: Professor P thinks that X and Y will be significantly related to each other, and I do not. Professor P collects some data bearing on this expectation and finds a significant relationship between X and Y—and touts this as inconsistent with my viewpoint.

But, being a good scientist, Professor P shares the data set with me. I proceed to conduct a great many significance tests on that data set and, following the general principle, adjust the alpha level accordingly. Lo and behold, the relationship between X and Y becomes nonsignificant, thereby confirming my theory and preventing Professor P from offering a substantive interpretation of the effect.

Surely this is an absurd consequence. The mere fact of my conducting additional significance tests on Professor P's data set should not be permitted to undermine the original conclusions. Yet, if one is committed to the principle that alpha should be adjusted to reflect the number of tests performed, then one will be committed to believing that my conduct—however mean-spirited and underhanded—was nevertheless statistically entirely justified. (And one will be compelled to conclude that, on the substantive question of the relationship of X and Y, my views should prevail over those of Professor P.)

To sharpen this example, imagine instead that Professor P had initially undertaken that large number of significance tests. If one supposes that, in such a circumstance, Professor P would be required to adjust the alpha level to avoid capitalizing on chance, then one is also committed to believing that my subsequent adjustment of the alpha level is not merely legitimate but actually required.

Parallel absurdities arise when researchers reanalyze their own data. In principle, at least, if I return to an old data set and conduct additional significance tests, I am presumably required to adjust the alpha level correspondingly. This, in turn, means that in some cases I will be required to submit a correction notice to the journal where the earlier data analyses were published—a notice indicating that some previously significant effects can no longer be treated as significant, because I conducted additional new significance tests.

These examples of data reanalysis, whether of one's own or someone else's data, have a clear implication: If the alleged principle were to be consistently applied, no research result could be deemed remotely secure. Any result could always instantly be undercut simply through the performance of additional significance tests completely unrelated to the finding of interest. Notice, for example, that my reanalysis of Professor

P's data need not have involved variables X and Y at all. The additional significance tests I computed could have involved only *other* variables in that data set. It is a bizarre principle indeed that says that one's beliefs about the relationship of X and Y should vary depending on the number of—not the results of but simply the sheer number of—other statistical tests that do not involve either X or Y.

As another illustration of the absurdities attendant to consistent application of the principle, consider the plight of the prolific researcher—the researcher who performs many significance tests. If, as the principle has it, alpha needs to be adjusted to avoid capitalizing on chance whenever a large number of significance tests are performed, then presumably the research community should insist on *careerwise* alpha adjustment. After all, the danger of capitalizing on chance arises whenever many tests are conducted, no matter whether the tests are conducted in a single study or spread across many studies. So when a researcher performs many significance tests, surely the putative principle is relevant ("as the number of significance tests increases, the risk of Type I error increases, and so the alpha level must be adjusted correspondingly"). It would be bizarre, of course, for the research community to say to a prolific researcher, "You are doing so many significance tests that we will require a more stringent alpha level for you than we do for other researchers." Yet, if the principle were consistently applied, that policy would have to be followed. (Of course, if the community did insist on such a policy, then prolific researchers would resort to the device of asking "virgin" investigators—people who had never before conducted a significance test—to perform tests for them.)

Or consider the circumstance of a research assistant who is asked to test the relationship between two variables in a data set, but who computes the matrix of correlations for all the variables in the data set and subsequently picks out the one of interest. If the principle were to be invoked consistently, alpha would have to be adjusted because of the large number of tests that were performed. That is, consistent application of the putative principle could mean that otherwise-significant results would have to be put aside simply because the assistant was lazy or forgetful (or knew how to get the program to compute all the correlations, but not how to obtain only the one of interest). It is surely absurd to be committed to believing that one's substantive conclusions about the relationship of variable 23 and variable 24 depend on whether "CORRELATIONS ALL" or "CORRELATIONS VAR23 VAR24" was typed as the SPSS command—and yet consistent adherence to the putative principle would indeed commit one to such beliefs.

To take stock of the argument thus far: The principle used to justify the use or imposition of these alpha-adjustment procedures is not applied consistently—and if applied in a consistent way would have bizarre consequences. Moreover, alpha-adjustment procedures have the undesirable

consequence of reducing statistical power. On the face of things, then, the policy of requiring or employing alpha adjustment as a consequence of the number of significance tests undertaken should be abandoned. No researcher should ever reduce, or be compelled to reduce, alpha because of the number of significance tests performed.[2]

The Root of the Problem

One way of describing the mistake involved in using or imposing alpha-adjustment procedures is to say that such procedures inappropriately localize Type I error. Alpha-adjustment procedures do localize Type I error in the sense that they mark out a particular set of tests (the "family" of tests) for which alpha adjustment is required. But there is no justifiable way of marking out a particular set of tests over which alpha adjustment is mandatory or appropriate; that is, there is no justifiable way to localize "capitalizing on chance."

For example, one cannot justifiably say that alpha must be adjusted when the tests are performed on one data set but not when the tests are spread across data sets. A researcher who performs 100 significance tests in 100 different studies has (ceteris paribus) the same overall Type I error risk as a researcher who performs 100 significance tests in one study.[3] Type I error risk inflation is a function of the number of tests performed, not a function of the data sets over which the tests are performed.[4]

Similarly, one cannot justifiably say that alpha must be adjusted when one researcher performs all the tests but not when different researchers perform the tests. If 100 researchers each perform one significance test, the overall Type I error risk is (ceteris paribus) the same as if one researcher performs all 100 tests. Type I error risk inflation is a function of the number of tests performed, not a function of the identity of the person who performs the tests.

And similarly, one cannot justifiably say that alpha must be adjusted when all the tests are performed simultaneously but not when the tests are spread out over time. A researcher who performs one significance test a day for 100 days has (ceteris paribus) the same Type I error risk as a researcher who performs 100 significance tests all at once. Type I error risk inflation is a function of the number of tests performed, not a function of the time at which the tests are performed.

It might be noticed that, with respect to this matter of localization, the justifications for alpha-adjustment procedures are often stated less carefully than they might be. Consider this example: "Many tests run at once will probably produce some significant results by chance alone, even if all the null hypotheses are true" (Moore & McCabe, 1999, p. 481). Notice the phrase "run at once," which implies that Type I error risk inflation arises specifically when many tests are conducted simultaneously. This is

incorrect. The inflation arises when many tests are conducted, regardless of whether the tests are performed simultaneously or sequentially.

Similarly, consider these examples:

> the major problem resulting from the performance of a series of analytical comparisons on a set of data is the unpleasant fact that the more comparisons we conduct, the more type I errors we will make when the null hypothesis is true. (Keppel, 1991, p. 164)

and "cumulative type I error is present whenever two or more statistical tests are performed in the analysis of a single experiment" (Keppel, p. 177). Phrases such as "on a set of data" or "in the analysis of a single experiment" might underwrite the inference that Type I error risk inflation arises specifically when many tests are conducted on the same data set. But this is incorrect. The inflation arises when many tests are conducted, regardless of whether the tests are performed on one data set or on many different data sets.[5]

In short, use or imposition of these alpha-adjustment procedures seeks to localize Type I error, but there is no justifiable way to mandate such localization. Type I error risk is not intrinsically localized (or localizable) in any way: It is not intrinsically localizable by data set, by researcher, by time, or by any other way of marking out a collection of tests. Rather, Type I error risk is diffused across the research enterprise; it is just the cost of doing business.

So if, using conventional Type I error standards (.05 alpha), a researcher tests a number of true null hypotheses, then in the long run 5% of the time the researcher will incorrectly conclude that a given null hypothesis is false. That may be regrettable, but it is a built-in cost of doing significance tests. That cost cannot be avoided, but it also cannot be localized.

Thus, one way of expressing the current argument is to say that the policy of using or imposing alpha-adjustment procedures involves inappropriate localization of Type I error risk. (The policy inappropriately localizes a risk that is better understood as spread throughout the research enterprise.) Because there is no justifiable way to mandate localization of Type I error, alpha-adjustment procedures are not defensible as a routine element of research practice and should never be employed or required.

COMPLEXITIES AND CAUTIONS

Rationales for Alpha Adjustment

It should be emphasized that the argument here opposes the invocation of Type I error risk inflation (attendant to a large number of tests) as

a basis for alpha-adjustment procedures. In other words, the argument is that the number of tests performed is not a sound basis for warranting alpha adjustments. Thus, the claims offered here are that no researcher can ever justifiably be required to undertake alpha adjustments because of the number of tests conducted and that no researcher should ever adjust alpha because of the number of tests conducted. The reason is that invocation of Type I error risk inflation is an unsatisfactory justification for such alpha adjustments because the justification is not principled (that is, it is not applied consistently) and because invoking that justification commits one to various absurd beliefs and practices.

Of course, a researcher might in some circumstances choose a more stringent alpha criterion (more stringent than the usual .05) for reasons other than the number of tests conducted. For example, if one thinks that one's results might form a basis for some consequential decision (a public policy decision, a medical intervention, and so forth), then one might decide that, for example, a 1% Type I error rate is more appropriate. But such a decision should not be based on, and cannot be justified by, the number of statistical tests being conducted in a study.

Planned vs. Unplanned Comparisons

It might be thought that a defensible formulation of the use of alpha-adjustment procedures can be had by invoking a distinction between planned and unplanned (post hoc) comparisons. Specifically, it has sometimes been suggested that planned comparisons do not require alpha adjustments, but that unplanned comparisons require some attention to familywise error rates (e.g., Keppel & Zedeck, 1989, p. 178). The rationale appears to be that if the researcher chooses which tests to perform only after seeing the data, then any particular such test (e.g., a particular comparison of two cell means) might have been chosen because the researcher has already surmised that that comparison is likely to turn out significant. That is, the researcher is said to have "implicitly" conducted some significance tests already (by virtue of having seen the data) and hence such implicitly conducted tests must be taken into account by adjusting the alpha level.

Right off the bat, it should be acknowledged that the distinction between planned and unplanned comparisons is a bit artificial, in the sense that any competent investigator can always concoct some justification for examining a given comparison and hence can present that comparison as having been planned. As Iacobucci (2001, p. 7) noted, a researcher might "delineate a very long list of contrasts and claim them all as planned, for who among us is not clever enough to create some rationalization if need be." Steinfatt (1979, p. 371) offered the image of the researcher "chanting 'I plan thee, I plan thee, I plan thee' just before conducting each significance

test," thereby vitiating the distinction.[6] A proposal to treat planned and unplanned comparisons differently with respect to alpha-adjustment procedures, then, actually offers little genuine protection against Type I error risk inflation.

However, even putting aside the artificiality and unenforceability of the distinction, it is not plain that differential treatment of planned and unplanned comparisons is appropriate. To put the matter most directly, differential treatment of planned and unplanned comparisons commits one to the curious belief that different conclusions should be drawn from a data set depending on what thoughts the researcher was having beforehand.

Imagine two researchers with identical data sets. (This is not necessarily entirely hypothetical; such a circumstance can arise when publicly available data sets are analyzed.) Given that data set A and data set B are identical, the same conclusions should be drawn from them. If, however, differential treatment of planned and unplanned comparisons is invoked, then researcher A and researcher B will be permitted very different conclusions if researcher A's comparisons were planned and researcher B's were unplanned. That is, differential treatment of planned and unplanned comparisons would have the researcher's previous thoughts, not simply the features of the data, influence the conclusions. This would be at least a strange methodological stance.[7]

Indeed, enforcing differential treatment of planned and unplanned comparisons transforms statistical analysis into something like a religious purity test. If researcher A's comparisons were planned then, by virtue of having those pure thoughts, researcher A is permitted to enter the realm of undiminished .05 alpha; by contrast, if researcher B's comparisons were unplanned then, because of having those impure sinful thoughts, researcher B is required to undertake alpha adjustments.

Perhaps part of the problem here is the implicit supposition that capitalizing on chance can occur only when comparisons are unplanned. This is incorrect. Even planned comparisons can capitalize on chance, and indeed inevitably do so. If one conducts 100 planned significance tests with .05 alpha on data where the null hypothesis is true, one will incorrectly reject the null five times in the long run. This capitalization on chance, however, is simply the risk of doing significance tests, no matter whether the tests in question are planned or unplanned.

Statistical Dragnets and Data Snooping

At least part of the rationale for imposing alpha adjustments, especially in the case of unplanned comparisons, appears to be an interest in discouraging the use of statistical dragnets (in which a large number of variables are examined and something less than a systematic rationale

underlies the inclusion of the variables) and data snooping (nosing around to locate some significant effects to report). The fear, obviously enough, is that investigators may be capitalizing on chance, because if enough tests are conducted, some significant effects will appear if only by chance.

One may certainly object to the inelegance and inefficiency of the dragnet as a research method and may quite appropriately urge investigators to pursue focused and specific research questions. But such judgments cannot legitimately be underwritten by concerns about Type I error risk inflation and cannot legitimately be enforced through the imposition of alpha-adjustment procedures. The Type I error risk incurred by an investigator who conducts 100 significance tests in statistical trawling and the Type I error risk incurred by the investigator who conducts 100 significance tests of narrowly focused hypotheses are (ceteris paribus) identical. Those who want to discourage statistical dragnets and data snooping need to find a good argument; the invocation of Type I error rate inflation is not such an argument.[8]

Null Hypothesis Significance Testing

One might think that the present argument applies specifically to null hypothesis significance testing (NHST) methods. Indeed, given recent critical discussions of NHST (for a useful collection, see Harlow, Mulaik, & Steiger, 1997), it might be thought that these arguments provide yet another reason for doubting the adequacy of NHST methods and for preferring examination of effect sizes and confidence intervals. This is not correct. Concerns about accumulated error are not limited to NHST methods, and the reasoning that leads to imposition of alpha-adjustment procedures for NHST can also be used to underwrite parallel adjustments for confidence intervals.

Concerns about accumulated statistical error will inevitably arise in any sample-based research enterprise, even without the use of NHST procedures, because of the conjunction of (a) the uncertainty associated with any sample result and (b) the examination of multiple uncertain results. Consider, for example, that when a confidence interval (CI) is constructed, the actual population value will sometimes lie outside the specified interval. For instance, in a hundred 95% CIs there will be, in the long run, five errors—that is, five cases in which the true value falls outside the interval. As the number of CIs constructed increases, the chances increase that at least one of them will fail to contain the population value. Given a large number of CIs, then, error risk inflation is a potential concern.

In fact, though it appears not widely appreciated, methods are available for adjusting the nominal (per interval) confidence level so as to take into account the number of CIs being constructed—precisely parallel to adjusting the nominal (per comparison) alpha level to take into account

the number of significance tests being conducted (see, e.g., Ramachandran, 1956; Roy & Bose, 1953). For example, a Bonferroni correction can be undertaken, so that, for instance, where five CIs are to be constructed and an overall error rate of 5% is desired, the investigator constructs 99% CIs rather than the usual 95% CIs (see, e.g., Sim & Reid, 1999, p. 192).

Thus the issues raised here, although expressed in terms of Type I error risk inflation and alpha-adjustment procedures, are actually not NHST-specific. The same defective reasoning that has been used to impose more stringent alpha criteria for multiple significance tests could be, and has been, invoked to impose more stringent criteria for multiple CIs,[9] and the objections offered here to alpha-adjustment procedures apply, *mutatis mutandis*, to such CI-adjustment procedures. Just as there is no way to justifiably mandate localization of Type I error risk, so there is no way to justifiably mandate localization of CI error risk. For example, a researcher who constructs one CI in each of 100 different studies has (ceteris paribus) the same cumulative CI error risk as a researcher who constructs 100 CIs in one study, and hence it would be unjustified to impose CI-adjustment procedures on the latter researcher but not the former.

That is to say, the issues that arise here cannot somehow be evaded by abandoning NHST procedures and instead pursuing analyses of effect sizes and confidence intervals. The same concerns about capitalizing on chance, data snooping, and the like can all arise—and can all lead to the inappropriate invocation of error-adjustment procedures—even outside the context of traditional NHST practices.

Contextualizing Statistical Tests

One might naturally have some concern about accepting the conclusions offered here. After all, there are some circumstances in which Type I error risk inflation would seem to cry out to be corrected. Imagine, for example, a circumstance in which an investigator reports 20 significance tests, of which only one is significant at .05 alpha. The nearly irresistible impulse is to suppose that this result may represent capitalization on chance and hence to conclude that this analysis requires alpha adjustment.

My argument is that such an impulse is in fact misplaced. There is nothing wrong with conducting 20 significance tests with .05 alpha, finding only one significant, and interpreting that one straightforwardly. Our discomfort with this circumstance is, I am arguing, a function of the particular context in which results are presented—a context that invites inappropriate localization of Type I error.

As a way of displaying this point, consider three cases, all involving independent tests with .05 alpha. In Case A, an investigator reports one study containing 20 significance tests and finds 1 significant. In Case B,

an investigator reports two studies: Study #1 reports 19 nonsignificant tests; Study #2 reports only 1 result, and it is significant. In Case C, an investigator reports 20 separate studies, each of which contains 1 significance test; in Studies #1–19 the result is nonsignificant, but Study #20 reports a significant result.

All three cases involve 20 tests and only 1 significant effect, and so (ceteris paribus) the overall Type I error rate is identical in the three cases. But Case A seems especially to set off intuitive alarm bells about the possibility of capitalizing on chance. In fact, anyone who believes that Case A requires alpha adjustment to avoid capitalizing on chance given the number of tests conducted is committed to believing that the same is true in Case B and Case C (same number of tests, same overall error rate, same likelihood of capitalizing on chance, same rationale for adjustment). And one is committed to such a belief even if the two studies in case B are reported in articles published in two different journals, even if the 20 reports in case C are spread out in time, even if the 20 reports in case C were to come from 20 different investigators, and so on. That is to say, anyone who worries about capitalization on chance when 20 tests are reported in a single article should have identical worries whenever 20 tests are reported. For example, if a journal issue contains four studies with five significance tests each, this concern about capitalization on chance should insist on what amounts to issuewise alpha adjustment.

Somehow readers are discomfited by seeing a large number of statistical tests in one study and so become agitated about the possibility of Type I error inflation even though they are not disturbed when an equally large number of tests—with equally large Type I error risk—is reported in a different format or context (e.g., spread out across a number of articles or spread out over a research career). Plainly, the difference in reaction is a consequence of how reported tests are contextualized.

So perhaps it will be useful to remember that any single test result, even if reported in isolation, can always be seen as part of some larger set of tests—other tests described in the same research report, other tests conducted by the same investigator in other studies, subsequent tests conducted over the same data set, other tests conducted by other investigators in other research projects, and so forth. Thus, even when one and only one test result is reported—and it is significant—this does not mean that capitalization on chance could not have happened. That one result is part of an indefinitely large number of tests, and thinking of it against the backdrop of those other tests may help to make clear that, even though reported in isolation, that result may nevertheless represent a chance effect.

Thus the difference between the case of 20 tests in 1 article where only 1 is significant and the case of 20 tests in 20 different articles where only 1 is significant is simply a difference in the context in which the

1 significant result is being viewed. When that 1 significant effect is viewed simultaneously with 19 other tests in a single article, capitalization on chance naturally suggests itself, but when that 1 significant result is the only result reported, then somehow chance seems an unlikely explanation. In fact, capitalization on chance is (ceteris paribus) equally likely in the two cases.

To come at this same matter from a slightly different direction: Believing that alpha-adjustment procedures are justified by Type I error risk inflation commits one to bizarre beliefs and defective research policies about issuewise and careerwise alpha adjustment, the consequences of data reanalysis, and so forth. There are two possible ways of avoiding being committed to these consequences. One way is the view being urged here, to abandon the idea that the use or imposition of alpha-adjustment procedures is warranted by the number of tests being conducted.

The other way of avoiding these unhappy consequences is to find a way of justifying some applications of alpha-adjustment procedures (e.g., the 20 tests in a single article) without being committed to other bizarre applications (e.g., the 20 tests in a single journal issue). Such a justification would necessarily have to involve underwriting the localization of Type I error, however. Expressed in terms of the earlier example, such a justification would require showing why Type I error is appropriately localized in one set of 20 tests (when the 20 tests occur in one article) but not in another set of 20 tests (when the 20 tests occur in one journal issue). And part of what has been argued here is that there is no justifiable way to mandate localization of Type I error: Type I error cannot be intrinsically localized by place of publication, by data set, by investigator, by time, and so forth. Any collection of 20 tests (ceteris paribus) has the same overall Type I error rate as any other collection of 20 tests, and hence it is unjustifiable to assert that the overall Type I error rate mandates or warrants alpha adjustment in some such collections but not others.

CONCLUSION

Type I error is a risk undertaken whenever significance tests are conducted, and the chances of committing a Type I error increase as the number of significance tests increases. But adjusting the alpha level because of the number of tests conducted in a given study has no principled basis, commits one to absurd beliefs and policies, and reduces statistical power. The practice of requiring or employing such adjustments should be abandoned.

NOTES

1. For a convenient summary treatment of these, see Keppel (1991, pp. 167–177). For examples of others, see Hochberg (1988), Holm (1979), Hommel (1988), and Rom (1990). For some discussion of these, see Shaffer (1995).

2. Some of the present arguments have been at least sketched before, though few commentators have seemed willing to draw what seems to me to be the natural conclusion, namely, that alpha adjustments should be abandoned. For instance, some who have acknowledged some absurdities attendant to alpha-adjustment procedures have nevertheless proposed (see Duncan, 1955, p. 16) or suggested modifications of (see the per-hypothesis approach of Wilson, 1962) such procedures. Even those who seem to want to conclude that alpha-adjustment procedures should be abandoned appear unable to bring themselves to do so wholeheartedly. Saville (1990), for example, argues against the use of various common adjustment procedures, but then seems to propose that tests with unadjusted alpha can only generate, not actually test, hypotheses.

3. Obviously, where *ceteris* is not *paribus*, all bets are off. See, for example, Proschan and Follmann (1995).

4. This formulation passes over a complexity. The exact familywise (FW) error rate varies depending on whether the comparisons involved are orthogonal or nonorthogonal, but "it is still accurate to say that the FW error rate increases with the number of comparisons we conduct regardless of orthogonality" (Keppel, 1991, p. 165). So, for instance, 100 independent significance tests conducted in 100 different studies yield the same familywise error rate as 100 independent significance tests conducted in a single study.

5. As indicated in note 4, the amount of Type I error risk inflation varies depending on whether the tests are independent, but Type I error risk increases as the number of tests increases regardless of the independence of the tests.

6. Steinfatt (1979) offers this image in an effort to show that because the distinction between planned and unplanned comparisons is easily undermined, attention to alpha adjustment issues is necessary for both planned and unplanned comparisons. The argument here, of course, is that because the distinction can so easily be undermined, neither sort of comparison should be subject to alpha adjustment.

7. Although the matter cannot be explored in detail here, it may be worth noticing that a methodological stance that treats planned and unplanned comparisons differently seems curiously inattentive to the needs of—and indeed to the very existence of—the larger research community. When researchers A and B offer different substantive conclusions by virtue of having had different alpha levels, what are others in the research community to believe? Do these others each flip a coin to decide between those conclusions? Does each try to imagine "if I were the researcher, would I have planned that comparison? If so, then I should adopt A's conclusion, but if not, then I'll adopt B's"? Obviously, one needs a decision procedure for persons other than the original researcher. Put differently, the research community at large needs to have a way of deciding what to believe, and it seems bizarre to employ a decision procedure in which the community's beliefs are held hostage to the original investigator's foresight. To express the concern here more generally: Handling planned and unplanned comparisons differently treats the research enterprise as fundamentally involving the beliefs of an isolated individual investigator (in the sense that the individual investigator's prior beliefs shape the conclusions to be drawn from the data and in the sense that the key question seems to be what conclusion the original investigator is to reach given the data). This seems manifestly incompatible with an image of scientific research as a community enterprise involving public argument and belief about shared evidence. The answer to the question, "What should the research community believe, given this data?", presumably should be independent of whether the original investigator did or did not have certain kinds of thoughts in advance of undertaking the statistical analysis.

8. In fact, it's not at all clear that one should discourage data snooping. After all, the data do not know what the researcher is thinking. If the researcher's data set contains a correlation of .53 between variable X and variable Y before the researcher begins to analyze the data set, then it contains that correlation, no matter what the researcher does or does not think or do. It does not matter whether the researcher thought about the correlation between X and Y previously, it does not matter whether the researcher ever examines the correlation between X and Y, and it does not matter whether the researcher planned to examine the correlation between X and Y. In short, it does not matter what the researcher was thinking—the correlation is .53 regardless. Hence our statistical treatment of the correlation should be the same no matter whether the correlation is discovered because the researcher specifically wanted to examine that correlation or because the researcher was snooping around the data set: The data are identical in the two cases and so the same conclusions should be drawn from the data. The researcher's poor rationale for including variables in the study, befuddled state of mind when planning the statistical analyses, and haphazard means of exploring the data do not alter the data, and so ought not alter the conclusions the research community draws from the data.

9. One is put in mind of Abelson's (1997, p. 13) prediction that "under the Law of Diffusion of Idiocy, every foolish application of significance testing will beget a corresponding foolish practice for confidence limits."

REFERENCES

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*, 12–15.

Duncan, D. B. (1955). Multiple range and multiple *F* tests. *Biometrika, 11*, 1–42.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Holt, Rinehart, & Winston.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.

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

Iacobucci, D. (2001). Can I test for simple effects in the presence of an insignificant interaction? *Journal of Consumer Psychology, 10*, 5–9.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches*. New York: W. H. Freeman.

Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

Moore, D. S., & McCabe, G. P. (1999). *Introduction to the practice of statistics* (3rd ed.). New York: W. H. Freeman.

Proschan, M. A., & Follmann, D. A. (1995). Multiple comparisons with control in a single experiment versus separate experiments: Why do we feel differently? *American Statistician, 49*, 144–149.

Ramachandran, K. V. (1956). Contributions to simultaneous confidence interval estimation. *Biometrics, 12*, 51–56.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika, 77,* 663–665.
Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.
Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* New York: Cambridge University Press.
Roy, S. N., & Bose, R. C. (1953). Simultaneous confidence interval estimation. *Annals of Mathematical Statistics, 24*, 513–536.
Saville, D. J. (1990). Multiple comparison procedures: The practical solution. *American Statistician, 44*, 174–180.
Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology, 46*, 561–584.
Sim, J., & Reid, N. (1999). Statistical inference by confidence intervals: Issues of interpretation and utilization. *Physical Therapy, 79*, 186–195.
Steinfatt, T. M. (1979). The alpha percentage and experimentwise error rates in communication research. *Human Communication Research, 5*, 366–374.
Wilson, W. (1962). A note on the inconsistency inherent in the necessity to perform multiple comparisons. *Psychological Bulletin, 59*, 296–300.