COMMUNICATION METHODS AND MEASURES, *1*(4), 267–273

It Takes a Family—a Well-Defined Family—to Underwrite Familywise Corrections

Daniel J. O'Keefe
Northwestern University

The false discovery rate (FDR) control procedures recommended by Matsunaga require the identification of a family of tests over which the procedures are applied. It is argued that Matsunaga's basis for grouping tests—that all the tests within the same null should be treated as a family, so long as there is a reasoning chain underlying the hypothesis—will, if applied in a principled (consistent) fashion, require bizarre and undesirable research practices. The underlying source of these difficulties appears to be an implicit (and unrealistic) focus on an isolated researcher conducting a single study, as opposed to a community of researchers with many studies and many tests.

Matsunaga's (this issue) proposals for handling concerns about Type I error are distressingly unclear, but it seems that he is arguing that (a) false discovery rate (FDR) control procedures should be preferred over procedures such as Bonferroni, but used (b) *only* for multiple tests of a given null hypothesis (not for tests of multiple different nulls) and (c) *only* when the hypothesis in question is supported by an appropriate chain of reasoning. I discuss each of these in turn. As will be seen, the key issue continues to be whether one can identify a principled basis on which to mark out a collection of tests as one over which alpha adjustment is required, without having consistent application of that principle lead to unhappy consequences. I believe Matsunaga's proposals do not provide any such basis.

Correspondence should be addressed to Professor Daniel J. O'Keefe, Department of Communication Studies, Frances Searle Building, Northwestern University, 2240 Campus Drive, Evanston, IL 60208-3545. E-mail: d-okeefe@northwestern.edu

FDR CONTROL PROCEDURES

FDR control procedures offer an interesting new approach to the concerns underlying the more familiar Bonferroni procedures. As Matsunaga indicates, FDR control procedures are designed to control the proportion of incorrect rejections of the null rather than being designed to minimize the chance of making even one incorrect rejection of a null (as with Bonferroni procedures). But whatever the advantages of FDR control procedures over practices such as Bonferroni corrections, these procedures still require the identification of a family (group) of tests over which the procedures are applied; hence, unless there is some suitable principled way of identifying a family of tests, FDR procedures will founder on the same shoals as do more familiar correction procedures.[1]

ONE NULL HYPOTHESIS

Matsunaga's proposal begins with the idea that "all statistical tests involved within the same H_O should be grouped as one 'family'" (2007, p. 251). That is, a set of tests' being "involved within the same H_O" is a necessary (and perhaps sufficient) condition for those tests being grouped as a family. But Matsunaga offers no apparent principled way to identify what counts as "the same" (i.e., just one) null hypothesis. (In a sense, the problem of defining in a principled way what counts as a family has been recast as the problem of defining in a principled way what counts as a single null hypothesis.)

To see the difficulties here, consider the following sort of case. Three researchers (in succession) independently test the null hypothesis that there is no difference in the relationship of exposure to media violence and aggressiveness across a particular set of four groups or conditions (e.g., young children, adolescents, younger adults, and older adults). The question is: Does this circumstance represent three different nulls or three tests of the same null? (Matsunaga's proposal requires having a principled answer to this question. To unambiguously identify a family of tests over which adjustment is properly applied, his proposal requires specification of what is to count as the "same" null, so questions such as these have to be addressed.) The problem is that under Matsunaga's proposal, either option commits one to absurd beliefs.

To treat these as three different nulls is to assert that these three studies actually test *different* hypotheses. Thus this option would commit one to

[1]Matsunaga's language often contrasts "FDR" and "familywise" (or "classic familywise") procedures, but this should not obscure the fact that even FDR procedures start with the identification of a family (collection, group) of tests to which the procedures are applied.

believing that generalization across studies is not possible because different studies of a given question, even if they appear to be testing the same null, are in fact (somehow) testing different nulls. So this option claims that it is improper to ever say that (for example) two studies support the same hypothesis (claim), or that one study supports a specific claim (hypothesis) but another does not support that claim—because no two studies ever actually test the same hypothesis. On this line of reasoning, replication or accumulation of research findings across studies is impossible, because each different study actually assesses a different null. Obviously, no one will want to be committed to this belief.

Consider the other option, that these three studies are studies of the same null. This option seems the sensible one, but if one accepts Matsunaga's proposal, there are unpalatable consequences. If "all statistical tests involved within the same H_O should be grouped as one 'family'" then these three studies would have to be grouped as one family. So Matsunaga's proposal would require that in undertaking significance tests, the second researcher must make a family of tests that includes the tests performed by the first researcher (after all, these are all tests of the same null), with corresponding re-computation of the first researcher's tests and, of course, consequences for the second researcher's tests (which now may have different results than if they had been analyzed on their own). When the third researcher comes on the scene, everyone re-computes again. And so on.

It gets worse. Matsunaga emphasizes restricting corrections to tests "within" the same null and thus eschews corrections when the null is $\mu_1=\mu_2$ ("only one comparison . . . does not require alpha adjustment" (2007, p. 257) but requires corrections when the null is $\mu_1=\mu_2=\mu_3$. As suggested, when multiple investigators test a null such as $\mu_1=\mu_2=\mu_3$, Matsunaga is presumably committed to requiring corrections that take into account the existence of other tests of that same null. But Matsunaga's reasoning appears also to require such corrections when multiple investigators each test a *simpler* null such as $\mu_1=\mu_2$, because (ex hypothesi) the researchers are testing "the same" null and hence "there is more chance for the null hypothesis to be falsely rejected" (2007, p. 253). The reader is invited to think of some well-studied hypothesis for which a great many studies have provided evidence—and to contemplate that, according to Matsunaga's proposal, all of the statistical tests accumulated over the years must now all be grouped as one family (because they all tested the same null), with attendant re-computation and corresponding re-assessment of our substantive beliefs about the phenomenon.

Now Matsunaga resists such conclusions, asserting that "even the most conservative approaches [to familywise error] do not reflect studies conducted by other investigators in other research projects" (2007, p. 254). But, this misses the point. The point is that the rationale Matsunaga offers (and, for that matter,

the rationales commonly invoked by others) to justify correction *requires* taking such other studies into consideration—otherwise the adjustment is not genuinely principled. It is capricious to treat differently (a) a case in which one investigator collects data to test the null hypothesis that $\mu_1=\mu_2=\mu_3$ and (b) a case in which two investigators separately collect data to test the null hypothesis that $\mu_1=\mu_2=\mu_3$. If Matsunaga's proposal mandates treating the first set of tests as a family over which adjustment is required, then a fortiori the second set must also be treated as such a family.[2]

Matsunaga is plainly discomfited by this prospect, as indicated by his desire to localize Type I error in a given study and thereby be given permission to ignore the conduct of other researchers. But brief invocations of Steinfatt's (1979) assertions that "the experiment" is a "natural unit" (2007, p. 251) will not suffice here. To see why, imagine an initial experiment that implicitly contains only one level of an independent variable (e.g., a persuasion effects study in which the communicator is high in credibility). A second experiment is conducted, otherwise identical to the first, but with other levels of that independent variable (to continue the example, the second experiment uses a low-credibility communicator). Obviously these data could be analyzed as a single study (in the example, a single study with credibility as an independent variable), but the question is: Does this circumstance represent "two experiments" or two halves of "one experiment?" Those who assert that the experiment is a natural unit are required to give an answer that employs a principled way of identifying the boundaries of "an experiment," but it is difficult to envision anything other than an arbitrary criterion. (And never mind complications such as "what if different investigators conduct the two parts?" or "what if the two parts are undertaken five years apart—or five months apart or five days apart?")

To be sure, it's awkward to confront questions such as "What is to be done when other researchers examine the same null?" But when several researchers are all testing the same null hypothesis, then if Matsunaga is to apply his standards in a principled way ("all statistical tests involved within the same H_O should be grouped as one 'family'") he is committed to grouping all those tests as a family over which adjustment is required.

---

[2]It is also capricious to treat differently (a) a case in which one investigator collects data to test the null hypothesis that $r_1=r_2=r_3$ and (b) a case in which two investigators separately collect data to test the null hypothesis that $r_1=r_2=r_3$; that is, it doesn't matter whether it's equivalencies of means or of correlations that's at issue. And it is capricious to treat differently (a) a case in which one investigator collects data to test the null hypothesis that $\mu_1=\mu_2$ (or that $r_1=r_2$) and (b) a case in which two investigators separately collect data to test the null hypothesis that $\mu_1=\mu_2$ (or that $r_1=r_2$). If all the tests relevant to a given null should be grouped as a family, then *all* the tests relevant to that null should be grouped as a family.

A CHAIN OF REASONING

Matsunaga appears to offer a second necessary condition to be met before tests are appropriately grouped as a family: "To the extent that a given research question or hypothesis is derived from a chain of theoretical argument and logical reasoning, its H_O provides a meaningful ground upon which Type I error can be localized and thereby controlled" (2007, p. 252). Taken at face value, this statement suggests that to the extent a given hypothesis is *not* so derived, to that same extent Type I error cannot be meaningfully localized. This seems to indicate that the hypothesis must be derived from a chain of theoretical argument and logical reasoning in order for adjustment to be appropriate (and indeed required). So, for example, concerning a null such as $\mu_1=\mu_2=\mu_3$, if I have no special reason to suppose that one or more of those equalities do not hold, then I should use ordinary uncorrected alpha levels; there is no reasoning chain and hence no basis for grouping the tests. By contrast, if I have "a chain of theoretical argument and logical reasoning" that leads me to think that one or more of those equalities do not hold, then I must apply some adjustment procedure.

Given the widely acknowledged bias toward publication of statistically significant results, this policy would obviously encourage researchers to eschew theoretical apparatuses. "Do I have theoretical reasons for expecting differences among these means? Oh, no, not me. I don't have any theoretical ideas here at all. So I will enjoy ordinary unadjusted .05 alpha for my tests. You researchers who have theoretical frameworks—*you* have to adjust your statistical procedures, and it will be harder for you to find statistically significant effects. Me, I get more statistical power because I lack theoretical guidance."

This is not a recipe for research progress.

Moreover, this reasoning-chain requirement seems inattentive to the existence of other investigators. For instance, suppose that when I examine another investigator's research, I conclude that her "chain of theoretical argument and logical reasoning" is flawed. Matsunaga's proposal would endorse my discarding the original researcher's analyses and conclusions, because my examination of those nulls would not be accompanied by any chain of arguments. Or, alternatively, suppose the primary researcher had no reasoning chain of the required sort (and so did no adjustments), but I can construct such a chain; presumably I must undertake adjustment. Or suppose nobody had the required reasoning at the time of initial data analysis, but 10 years later such a theoretical rationale appears—must we now go back and recompute our tests?

Given that the data are what's of fundamental interest—regardless of what reasoning the investigator did or didn't have—surely it would be preferable for investigators to simply report unadjusted results; other members of the research community could then adjust (or not) depending on their own inclinations and

auxiliary reasoning. Certainly nothing in Matsunaga's argument *mandates* any alpha adjustment by the original investigator nor by any other member of the research community.

Indeed, Matsunaga as much as concedes that his proposal does not supply a justification for mandating corrections in any specific case: "There would be a certain degree of latitude as to how to frame a given H_O, as would any argumentation reflect the researcher's particular theoretical standpoint. On this front, any decision made on whether and how to adjust alpha would be debatable" (2007, p. 251). But if there is not yet any defensible and decisive general rationale for the imposition of alpha-adjustment procedures, then surely the practice of requiring or employing such adjustments is improper.

MULTIPLE INVESTIGATORS, MULTIPLE STUDIES, AND RESEARCH COMMUNITIES

"Alpha inflation" is a chimera and hence there is no need to "control" it. Type I errors are a cost of doing business in a sample-based research enterprise. The appearance of alpha "inflation" is created by focusing inappropriately on some specific set of tests, usually within a single study (for some elaboration of this idea, see O'Keefe, 2003, pp. 442–444). And this narrowed focus represents an unfortunate tendency (in thinking about Type I error) to imagine an isolated investigator undertaking just one study—and to ignore the existence of the larger research community (other people who need to decide what to believe), other studies (by the original investigator or by others), and other tests. For precisely that reason, any proposed principle for specifying a family of tests should be confronted with questions such as "what if other researchers collect new data concerning that hypothesis?" and "what if other researchers do not share the investigator's reasoning?" and "what if many different researchers run statistical tests over the same data set?" My belief is that the efforts to date at specifying what counts as a family of tests over which correction is mandated—whether "all the tests conducted over a given data set," "all the tests within a given article," "all the tests of a given theory," or "all the tests within a given null"—are, if applied in a principled way to the situations contemplated by such questions, committed to requiring bizarre and undesirable research practices that no sensible person would endorse. For that reason, those proposed family boundaries are unsatisfactory.

None of this means Type I error is unimportant. But the best way to address concerns about false discoveries is replication. In a sample-based research enterprise, mistakes are inevitable. But replication will sort matters out in a far more decisive way than can any possible fiddling with statistical procedures.

REFERENCES

Matsunaga, M. (2007). Familywise error in multiple comparisons: Disentangling a knot through a critique of O'Keefe's arguments against alpha adjustment. *Communication Methods and Measures, 1*, 243–265.

O'Keefe, D. J. (2003). Against familywise alpha adjustment. *Human Communication Research, 29*, 431–447.

Steinfatt, T. M. (1979). The alpha percentage and experimentwise error rates in communication research. *Human Communication Research, 5*, 366–374.