ACADEMIA | Letters

# *There's No Need to Lower the Significance Threshold When Conducting Single Tests of Multiple Individual Hypotheses*

Mark Rubin, The University of Newcastle, Australia

---

During null hypothesis significance testing, a multiple testing problem occurs when researchers test a *joint null hypothesis* using a *union-intersection testing* approach (e.g., Kim et al., 2004; Parker & Weir, 2020; Roy, 1953). A joint null hypothesis comprises two or more constituent null hypotheses that can each be subjected to separate significance tests. For example, consider the joint null hypothesis that "male participants have *neither* higher academic self-esteem scores *nor* higher social self-esteem scores than female participants." In this case, a researcher could conduct a significance test on each of the two constituent null hypotheses: (a) "male participants do not have higher *academic* self-esteem scores than female participants," and (b) "male participants do not have higher *social* self-esteem scores than female participants." During union-intersection testing, it is logical to reject the overall joint null hypothesis if *at least one* of its constituent null hypotheses is rejected. So, for example, it is logical to reject the joint null hypothesis that "male participants have *neither* higher academic self-esteem scores *nor* higher social self-esteem scores than female participants" if we reject the constituent null hypothesis that "male participants do not have higher academic self-esteem scores than female participants."

The multiple testing problem occurs because the greater the number of constituent null hypotheses that comprise the joint null hypothesis, the greater the number of opportunities that a researcher has to incorrectly reject *at least one* constituent null hypothesis during union-intersection testing, and so the greater the probability of incorrectly rejecting the overall joint null hypothesis. More formally, if a researcher uses a significance threshold (alpha level) of $\alpha$ for each of $k$ independent tests of their joint null hypothesis, then the Type I error rate

*Academia Letters, March 2021* 

**Corresponding Author:** Mark Rubin, Mark-Rubin@outlook.com
**Citation:** Rubin, M. (2021). There's No Need to Lower the Significance Threshold When Conducting Single Tests of Multiple Individual Hypotheses. *Academia Letters*, Article 610.
https://doi.org/10.20935/AL610.

for the joint null hypothesis will be $1 - (1 - \alpha)^k$. For example, if a researcher tests a joint null hypothesis that comprises two constituent null hypotheses, and they use a significance threshold of $p \leq .050$ for each of their two constituent significance tests, then the Type I error rate for rejecting the joint null hypothesis will be .098 (i.e., $1 - [1 - .050]^2$) rather than .050.

A typical solution to this multiple testing problem is to lower the significance threshold for each constituent test of the joint null hypothesis. For example, in the preceding case, the significance threshold for each constituent null hypothesis could be lowered from $p \leq .050$ to $p \leq .025$. This Bonferroni correction would maintain the significance threshold for the joint null hypothesis at .050. However, researchers sometimes become confused about when to apply this type of correction.

A common multiple testing myth is that it is necessary to lower the significance threshold when undertaking single tests of multiple individual null hypotheses *even when researchers are not testing a joint null hypothesis using a union-intersection approach* (cf. Matsunaga, 2007; O'Keefe, 2003). For example, it is often assumed that a researcher who wants to conduct significance tests of two individual null hypotheses using a significance threshold of $p \leq .050$ will have an inflated Type I error rate of .098 (i.e., $1 - [1 - .050]^2$) and, consequently, they will need to lower their significance threshold in order to compensate for this error inflation.

To be clear, it is correct that, assuming a significance threshold of $p \leq .050$, the probability of incorrectly rejecting at least one of two individual null hypotheses is .098 (i.e., $1 - [1 - .050]^2$). However, my key point is that *researchers should not be interested in this familywise error rate if they are testing each null hypothesis individually rather as part of a family of tests of a joint null hypothesis*. As one of the originators of modern hypothesis testing warned, "many errors in computing probabilities are committed because of losing sight of the set of objects to which a given probability is meant to refer" (Neyman, 1950, p. 15). In the current case, the error occurs because researchers lose sight of the fact that their nominal Type I error rate applies to two separate decisions about two individual null hypotheses rather than to a single decision about a joint null hypothesis. Hence, if researchers make separate decisions about each individual null hypothesis that they test, then the probability of incorrectly rejecting each null hypothesis due to random sampling and measurement error will be no higher than their significance threshold for each hypothesis, and no correction to the significance threshold is required.

To illustrate, consider a researcher who conducts a single test of the individual null hypothesis that "male participants do not have higher *academic* self-esteem scores than female participants." Now imagine that the researcher conducts an additional single test of the individual null hypothesis that "male participants do not have higher *social* self-esteem scores than female participants." In this case, it is not necessary for the researcher to lower their




**Corresponding Author:** Mark Rubin, Mark-Rubin@outlook.com
**Citation:** Rubin, M. (2021). There's No Need to Lower the Significance Threshold When Conducting Single Tests of Multiple Individual Hypotheses. *Academia Letters*, Article 610.
https://doi.org/10.20935/AL610.

significance threshold for either test as long as they make separate decisions and probability statements about each individual hypothesis (e.g., "male participants had higher *academic* self-esteem scores than female participants, $p$ = .031," and "male participants had higher *social* self-esteem scores than female participants, $p$ = .027"). A lowered significance threshold would only be necessary if the researcher made a decision and probability statement about a joint hypothesis (e.g., "male participants had higher *academic or social* self-esteem than female participants, $p$ = .015"). In short, there is no reason to adjust the significance threshold when testing two or more null hypotheses unless they comprise a joint null hypothesis that undergoes union-intersection testing (Cook & Farewell, 1996; Hurlbert & Lombardi, 2012; Matsunaga, 2007; Rubin, 2017, 2020, 2021; Savitz & Olshan, 1995, p. 906; Tukey, 1953).

Some researchers may feel reluctant to let go of the above multiple testing myth because they understand, correctly, that the more significance tests that they conduct, the more chance they will have of making a Type I error. However, this "more tests, more errors" concern needs to be interpreted in the context of the *number of hypothesis decisions* that are made. Obviously, the more decisions that a researcher makes, the more chance they will have of making a Type I error. However, this fact does not imply that their Type I error rate will be inflated *for any given decision*. In particular, there will be no Type I error rate inflation if the ratio of tests-to-decisions remains at 1, so that $k$ = 1 in the formula $1-(1-\alpha)^k$ for each decision. There will only be a Type I error rate inflation if multiple test results are used to make a single decision because, in this case, $k$ will be higher than 1 for that decision. For example, if a researcher uses 20 tests, each with a significance threshold of $p \leq .050$, to make a single decision about a single joint null hypothesis, then they will have a probability of .64 of making a Type I error due to their multiple testing of this null hypothesis (i.e., $1-[1-.050]^{20}$. However, if they use 20 tests, each with a significance threshold of $p \leq .050$, to make 20 separate decisions about each of 20 individual null hypotheses, then they will have a probability of .050 of making a Type I error in each case (i.e., $1-[1-.050]^{1}$). Hence, an adjustment to the significance threshold is only warranted in the former case and not in the latter case.

In summary, it is only necessary to lower the significance threshold when undertaking multiple tests of a single joint null hypothesis using an union-intersection approach. It is not necessary to lower the significance threshold when undertaking single tests of multiple individual null hypotheses.




**Corresponding Author:** Mark Rubin, Mark-Rubin@outlook.com
**Citation:** Rubin, M. (2021). There's No Need to Lower the Significance Threshold When Conducting Single Tests of Multiple Individual Hypotheses. *Academia Letters*, Article 610.
https://doi.org/10.20935/AL610.

## References


Cook, R. J., & Farewell, V. T. (1996). Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 159,* 93-110. https://doi.org/10.2307/2983471

Hurlbert, S. H., & Lombardi, C. M. (2012). Lopsided reasoning on lopsided tests and multiple comparisons. *Australian & New Zealand Journal of Statistics, 54,* 23-42. http://dx.doi.org/10.1111/j.1467-842X.2012.00652.x

Kim, K., Zakharkin, S. O., Loraine, A., & Allison, D. B. (2004). Picking the most likely candidates for further development: Novel intersection-union tests for addressing multi-component hypotheses in comparative genomics. *Proceedings of the American Statistical Association, ASA Section on ENAR Spring Meeting* (pp. 1396-1402). http://www.uab.edu/cngi/pdf/2004/JSM%202004%20-IUTs%20Kim%20et%20al.pdf

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. https://doi.org/10.1080/19312450701641409

Neyman, J. (1950). *First course in probability and statistics.* Henry Holt.

O'Keefe, D. J. (2003). Colloquy: Should familywise alpha be adjusted? *Human Communication Research, 29*, 431-447. https://doi.org/10.1111/j.1468-2958.2003.tb00846.x

Parker, R. A., & Weir, C. J. (2020). Non-adjustment for multiple testing in multi-arm trials of distinct treatments: Rationale and justification. *Clinical Trials,* 1-5. https://doi.org/10.1177/1740774520941419

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics, 24,* 220-238. https://doi.org/10.1214/aoms/1177729029

Rubin, M. (2017). Do *p* values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology, 21,* 269-275. http://dx.doi.org/10.1037/gpr0000123

Rubin, M. (2020). Does preregistration improve the credibility of research findings? *The Quantitative Methods for Psychology, 16*(4), 376–390. https://doi.org/10.20982/tqmp.16.4.p376


 

**Corresponding Author:** Mark Rubin, Mark-Rubin@outlook.com
**Citation:** Rubin, M. (2021). There's No Need to Lower the Significance Threshold When Conducting Single Tests of Multiple Individual Hypotheses. *Academia Letters*, Article 610.
`https://doi.org/10.20935/AL610`.

Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese*. https://doi.org/10.1007/s11229-021-03276-4

Savitz, D. A., & Olshan, A. F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology, 142*, 904-908.

Tukey, J. W. (1953). *The problem of multiple comparisons.* Princeton University.

 

**Corresponding Author:** Mark Rubin, Mark-Rubin@outlook.com
**Citation:** Rubin, M. (2021). There's No Need to Lower the Significance Threshold When Conducting Single Tests of Multiple Individual Hypotheses. *Academia Letters*, Article 610.
https://doi.org/10.20935/AL610.