Adjusting for Multiple Comparisons

The Office of Evaluation Sciences' standard approach to adjusting for multiple comparisons is to control the familywise error rate (FWER), which is the probability of seeing at least one false positive, or Type I error, in a "family" of statistical tests. The purpose of controlling the FWER is to mitigate the risk of accepting a statistical result that might be a false positive as support for a policy decision. In particular, our approach is designed to achieve an overall Type I error rate of 5% across all the hypotheses we test in support of a single policy decision.

When to Adjust for Multiple Comparisons

We make adjustments to control the FWER when:

- We have more than one hypothesis or outcome (family of hypotheses) upon which one policy decision will be based, and
- The hypotheses are declared confirmatory in the analysis plan

We do not control the FWER when:

- We do not have more than one hypothesis/outcome, or
- The hypotheses are declared exploratory in the analysis plan

What Adjustment Procedures to Use

When hypotheses or outcomes are uncorrelated, OES uses the Holm-Bonferroni procedure to control the FWER. When hypotheses or outcomes are correlated, we use simulation based on repeated randomization and hypothesis testing to control the FWER.

How to Report the Adjustment

To report this information, we report the unadjusted *p*-values and confidence intervals in the text of the abstract. We then present the adjusted per-test significance levels in a footnote for each reported *p*-value.

Example text to use in abstracts

For Holm-Bonferroni:

> The results estimate that individuals who received the postcard reminder in October received 0.27 percentage points more vaccinations than individuals in the control group ($p < .01$, 95% CI [0.12 pp, 0.42 pp]).[1]

Update June 2020

[1]This result is statistically significant at significance level .017, which controls the familywise error rate at .05 by using the Holm-Bonferroni procedure.

For simulations:

> The results estimate that individuals who received the postcard reminder in October received 0.27 percentage points more vaccinations than individuals in the control group ($p < .01$, 95% CI [0.12 pp, 0.42 pp]).[1]
>
> [1]This result is statistically significant at significance level .017, which controls the familywise error rate at .05 based on repeated simulations of randomization and hypothesis testing.

Example text to use in reports or presentations

For a technical report/presentation:

> At OES, we also report adjusted significance levels per test, the level at which we consider a *p*-value to be significant, when we report results for more than one hypothesis upon which a policy decision will be based. The adjusted significance levels per test control the familywise error rate at .05. The purpose of reporting adjusted significance levels per test is to mitigate the risk of accepting results for a policy decision that may be false positives. We calculate the adjusted significance level per test either through the Holm-Bonferroni procedure or through repeated simulations of randomization and hypothesis testing.

For a policy presentation:

> The more hypotheses we test for a policy decision, the greater the probability that we see a false positive. We control for the probability of seeing a false positive by making our significance level more conservative. We do this in such a way that we limit ourselves to an overall error rate of 5% across all the hypotheses we test in support of a single policy decision.