NBER WORKING PAPER SERIES

BEYOND BONFERRONI: MULTIPLE TESTING IN EMPIRICAL RESEARCH

Sebastian Calónico
Sebastian Galiani

Working Paper 34050
http://www.nber.org/papers/w34050

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2025, Revised June 2026

We thank Raul A. Sosa for excellent research assistance and ChatGPT 5.1 (OpenAI) for help with the editing of this paper. All remaining errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Beyond Bonferroni: Multiple Testing in Empirical Research
Sebastian Calónico and Sebastian Galiani
NBER Working Paper No. 34050
July 2025, Revised June 2026
JEL No. C1, C11, C13, C15, C18

ABSTRACT

Empirical work in economics routinely tests many hypotheses at once, but applied researchers often lack clear guidance on how to handle the resulting multiplicity. This paper offers a practical guide. We argue that the choice of error criterion — the chance of any false rejection, or the share of false rejections among discoveries—should follow from the structure of the decision the tests inform. Beyond standard corrections like Bonferroni and Holm, we make the case for resampling-based procedures, particularly Romano–Wolf, which exploit dependence among test statistics — the central feature of multiple testing in economic applications. We also recommend hierarchical methods that use causal or logical structure to deliver more powerful results. By linking these tools to real applications, we provide a clear roadmap, anchored in pre-specification, for credible empirical work.

Sebastian Calónico
University of California, Davis
Graduate School of Management
scalonico@gmail.com

Sebastian Galiani
University of Maryland, College Park
Department of Economics
and NBER
sgaliani@umd.edu

1 INTRODUCTION

Empirical research in economics routinely tests many hypotheses at once: across multiple outcomes, subgroups, and model specifications. But researchers often lack clear guidance on how to handle the resulting multiplicity in a way that supports credible inference. The inferential tools designed for a single test do not carry over unchanged. The probability of at least one false positive grows mechanically with the number of tests, even when each individual test meets its nominal Type I error rate. The findings — and the decisions built on them — are less reliable than the nominal error rate suggests.

Consider a job training program evaluated on eight labor-market outcomes across three demographic subgroups, giving $m = 24$ effect estimates. A simple rule is to adopt the program if at least one of the 24 tests rejects the null of no effect at the conventional five percent level. If the tests were independent, the decision maker would commit funds to a program with no effect on any outcome about 71 percent of the time, since

$$\mathbb{P}(\geq 1 \text{ false positive}) \;=\; 1-(1-\alpha)^m \;=\; 1-0.95^{24} \;\approx\; 0.71. \tag{1}$$

The tests are not independent. Earnings, employment, and hours co-move, and within each subgroup the same individuals contribute to every outcome. Dependence is the central feature of the problem in economics. It pulls the false-adoption rate below 0.71, but in any realistic correlation structure the rate remains well above five percent.

Despite the menu of methods now in use, applied economists face a recurring choice problem: which method fits which design, when to apply it, and how to interpret the results. Where genomics and high-dimensional inference work with thousands of tests and analytical bounds on worst-case dependence, applied economics typically works with families small enough that the dependence is estimable from the data. What has been missing is a framework that ties the choice of adjustment to the structure of the decision the tests inform.

The use of multiple-testing adjustments has been growing, steadily but unevenly, in two distinguishable patterns. First, the share of empirical papers in economics that report a formal multiple-testing adjustment, shown in Figure 1, has roughly quadrupled since 2010, with a marked acceleration from the mid-2010s onward. Second, Panel B shows that the rise is not driven by a single procedure: Bonferroni and Holm remain the workhorse but have leveled off, Benjamini–Hochberg has accelerated since the late 2010s, and Romano–Wolf, though still the smallest family, is the fastest-growing. What might once have been a single hypothesis test has, in many applications, become a portfolio of them.

This paper develops a conceptual framework for multiple testing in applied economics, recommends specific procedures for applied research, and supports those recommendations with simulation evidence.[1] The choice of error criterion follows from the structure of the decision the tests inform. The relevant family is the set of tests that jointly inform one decision. A single paper can contain several families when it addresses more than one.

In confirmatory settings, where any false positive can lead to a wrong decision, the natural criterion is the familywise error rate (FWER), and our recommended default is the Romano–Wolf bootstrap stepdown procedure (Romano and Wolf, 2005a,b, 2016), which exploits the dependence structure typical of economic data and tends to deliver higher power than the analytical alternatives. The Holm procedure (Holm, 1979) is a valid fallback when bootstrap validity is in doubt. In exploratory settings, where individual false leads are tolerable, a common alternative is the false discovery rate, for which the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) is the standard. When hypotheses come with a natural ordering, hierarchical testing concentrates the error budget on primary claims and raises overall power. Simultaneous confidence intervals, constructed from the same resampling distribution, complement the adjusted tests by addressing the magnitude question that point rejections leave open.

[1]The procedures recommended here are implemented in `mht`, a Stata command we developed as a companion to this paper (Calonico et al., 2026). The command, together with the data and code for the empirical application, is available at `https://www.sebastiangaliani.com/academic-code`.

None of this is credible without pre-specification of families and methods, ideally in a pre-analysis plan. Pre-specification disciplines which tests count, but on its own it does not control the error rate across them.

Section 2 shows how multiplicity in economics reduces to a small number of design patterns, all with dependent test statistics. Section 3 derives the FWER/FDR choice from a decision-theoretic loss minimization, and addresses the interpretation of an individual hypothesis when many are tested. Section 4 surveys the methods and lands the case for Romano–Wolf as the confirmatory default. Section 5 discusses pre-specification and its complement, publication-bias correction. Section 6 illustrates the framework end-to-end on the Piso Firme evaluation of Cattaneo et al. (2009). Section 7 concludes.

2 Multiple Testing in Applied Economics

Two features of multiple testing in applied economics shape the methodological choices that follow. First, multiplicity arises in a small number of recognizable patterns tied to the research design, so the relevant family of hypotheses is identifiable rather than open-ended. Second, the test statistics those designs produce are dependent in structured ways that can be estimated from the data, a regime in which the textbook corrections built for worst-case dependence give up power without compensating gains. Section 2.1 develops the patterns; Section 2.2 develops the dependence.

2.1 *A Taxonomy*

We identify five recurring patterns of multiplicity in applied economics.

(a) Multiple outcomes, same treatment. The most common form of multiplicity arises when a single treatment is evaluated on several outcome variables. Cattaneo et al. (2009) study the Piso Firme program, which replaced dirt floors with cement floors in poor Mexican households, examining a wide battery of outcomes spanning child health, adult wellbeing, and housing quality. Each outcome was tested separately for a treatment effect,

and the question of whether the program "works" aggregates across the individual tests. The Moving to Opportunity experiment (Kling et al., 2007) provides another canonical example: families receiving housing vouchers were evaluated on earnings, employment, physical health, mental health, risky behavior, and educational achievement. In both Piso Firme and MTO, the dependence among tests is strong and positive, because outcomes such as parasitic infestations and diarrhea reflect related dimensions of child health, and depression and life satisfaction measure overlapping aspects of adult wellbeing. This positive dependence has direct consequences for the choice of adjustment method (Section 2.2).

Index aggregation is one of the principal ways applied work handles multi-outcome multiplicity: outcomes that share an underlying construct are combined into a summary measure before testing (Kling et al., 2007; Anderson, 2008), reducing the family size at the cost of a coarser substantive claim. Whether to aggregate is a design decision tied to the substantive question, not a statistical one. Section 4.4 returns to indices in hierarchical procedures, and Section 4.6 gives operational guidance.

(b) Subgroup analyses. A second source is heterogeneous treatment effects across subgroups defined by demographic or baseline characteristics. A program evaluation that examines effects by gender, age, and education has implicitly tested many hypotheses, often exploratory: the effective family includes all subgroups examined, not just the ones reported. The dependence structure tracks the subgroup overlap. Mutually exclusive subgroups under a tying constraint (men versus women with a fixed overall mean) produce negatively correlated test statistics; disjoint subsamples without such a constraint yield approximately independent ones; overlapping subgroups ("young women" inside both "women" and "young people") yield positively correlated statistics because they share observations.

(c) Multiple specifications and robustness checks. A distinctive form of multiplicity arises when the same hypothesis is tested across different regression specifications, varying

the control variables, functional form, or sample restrictions. The dependence is typically very high: each regression uses the same (or nearly the same) dependent variable and treatment variable. When the added controls are orthogonal to the treatment, the Frisch–Waugh–Lovell theorem makes the coefficient identical across specifications, the t-statistics line up, and the family-wise error rate is approximately α regardless of K. A Bonferroni correction would impose a heavy power penalty for a multiplicity problem that does not exist. When controls correlate with the treatment, the coefficients shift across specifications and the dependence structure becomes nontrivial; whether the resulting tests constitute genuine multiplicity or sensitivity analysis depends on whether the specifications test different hypotheses or the same hypothesis under different modeling assumptions. In the framework of Viviano et al. (2026), the marginal cost of adding a robustness check is small once the underlying analysis is in place, so the optimal correction collapses toward none—a formal counterpart to the position that genuine sensitivity analyses are not multiplicity problems. When the specification is instead chosen by the data—as in lasso or stepwise selection—the relevant correction is not a multiplicity adjustment but post-selection inference, which conditions on the selection event (Lee et al., 2016).

(d) Multiple treatments or treatment arms. RCTs that compare several treatment arms against a common control group induce positive correlation among the treatment-control comparisons through the shared control. With K balanced arms and a control of equal size, the shared control mean $\bar{y}_c$ contributes half of the sampling variance of each contrast $\hat{\beta}_k = \bar{y}_k - \bar{y}_c$, so any pair of contrasts has correlation 0.5. The Moving to Opportunity experiment is again instructive (Kling et al., 2007): families were randomized into a traditional Section 8 voucher arm, an experimental voucher conditional on moving to a low-poverty neighborhood, and a control group, generating two treatment-vs-control comparisons that share a reference group. Each comparison is a separate test, and the conclusion about which intervention to scale up aggregates across them. Because the dependence is positive and identified by the design, Bonferroni-type corrections that

ignore the shared-control structure are conservative; resampling methods such as Romano–Wolf absorb the dependence and recover the lost power, the correction List et al. (2019) develop for multi-arm, multi-outcome experiments.

(e) Mixed and hierarchical structures. The patterns above frequently combine: a typical evaluation has a primary outcome family, a mechanism family that probes how the program operates, and an exploratory family of subgroups or downstream effects. When the substantive logic of the study orders these layers naturally, hierarchical testing procedures concentrate the error budget on the primary claims by testing higher-priority hypotheses first and proceeding down the hierarchy only when prior tests reject. Hierarchical adjustment is the most underused tool in the applied multiple-testing toolkit. Section 4.4 develops the menu of hierarchical procedures, and Section 6 quantifies the gains over flat methods in simulation.

2.2 *Dependence: The Central Feature*

Across all five patterns, the test statistics are dependent—typically positively—and the dependence is structured by the research design rather than worst-case. Of the features that distinguish multiple testing in economics from the textbook treatment, this is the most consequential and the least addressed. This section develops the consequences.

Positive dependence reduces the true FWER below the level implied by the independence formula $1 - (1 - \alpha)^m$. To see the exact mechanism, consider a stylized example in which a researcher estimates m regressions of different outcomes on the same binary treatment variable:

$$y_j = \mu_j + \beta_j x + \varepsilon_j, \qquad j = 1, \ldots, m. \tag{2}$$

Now suppose the outcomes are rescalings of one another: $y_j = a_j + b_j \cdot y_1$ for constants a_j and $b_j > 0$. This is an extreme case, but it captures the intuition that outcomes such as "monthly earnings" and "annual earnings" contain the same information. Substituting into (2) shows that the residuals satisfy $\varepsilon_j = b_j \varepsilon_1$ exactly, so both classical and

heteroskedasticity-robust standard errors scale by b_j. Under this assumption,

$$\hat{\beta}_j = b_j \hat{\beta}_1, \tag{3}$$

$$\mathrm{SE}(\hat{\beta}_j) = b_j \mathrm{SE}(\hat{\beta}_1), \tag{4}$$

$$t_j \;=\; \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} = \frac{b_j \hat{\beta}_1}{b_j \mathrm{SE}(\hat{\beta}_1)} \;=\; t_1 \quad \text{for all } j. \tag{5}$$

All m t-statistics are identical. Either every test rejects or none does. There is no possibility of a false rejection on one hypothesis but not another, and the FWER is simply α by construction. Now consider how different methods perform:

- *Bonferroni* compares each t-statistic to the critical value at level α/m. Because all t-statistics are the same, the actual FWER is α/m, far below the nominal target, and the researcher pays a considerable cost in power for a problem that, in this stylized case, does not exist. With $m = 20$, the effective significance level drops from 5% to 0.25%.
- *Holm's stepdown procedure* orders the p-values and compares the smallest to α/m, the same threshold as Bonferroni. When all p-values are identical, Holm suffers the same power loss.
- *Romano–Wolf* bootstraps the data and computes the joint distribution of the test statistics under the null. Because the bootstrap replicates the dependence structure, it recovers the fact that the maximum t-statistic is approximately equal to each individual t-statistic. The adjusted critical values are close to the unadjusted ones, and the procedure correctly controls FWER at approximately α without sacrificing power.

In practice, outcomes are never perfectly correlated, so the effective number of independent tests lies between 1 and m. The key insight is that the closer the outcomes are to perfect dependence, the larger the power gain from resampling methods relative to Bonferroni or Holm. Even moderate positive correlations among test statistics, of the

kind that arises naturally in multi-outcome program evaluations (e.g., Kling et al., 2007; Cattaneo et al., 2009), produce meaningful power gains from the Romano–Wolf procedure (Section 6).

Negative dependence is also possible and works in the opposite direction. The cleanest applied example arises in tests on categorical shares that partition the sample: a balance test comparing the share female in treatment versus control, run in parallel with a test comparing the share male, produces perfectly negatively correlated test statistics because the shares sum to one. In this regime, the independence formula $1 - (1 - \alpha)^m$ *overstates* the true FWER, and Bonferroni becomes conservative for a different reason than under positive dependence. Resampling methods adapt to the sign of dependence by estimating the joint distribution directly, a further advantage when mixed positive and negative dependencies appear in the same family.

These features set economics apart from the high-dimensional applications most multiple-testing methods were designed for: the standard treatises on large-scale inference (Efron, 2010) and pharmaceutical statistics (Dmitrienko et al., eds, 2009) address the large-m regime where dependence must be bounded analytically, but offer less guidance for the small-m, estimable-dependence regime of applied economics. For the applied economist, dependence is an opportunity for power gain, not a worst-case threat to control.

3 A Decision-Theoretic Framework

Three questions in multiple testing are usually answered by convention: which error rate to control, what level to control it at, and how a reader should interpret an individual p-value when many tests were run. This section grounds each in the underlying decision problem instead. Section 3.1 works through two examples — program adoption and mechanism screening — in which the error rate falls out of the action the test informs and the significance level falls out of the cost ratio, so the familiar 0.05 emerges as one calibration among many. Section 3.2 names and defines the criteria. Section 3.3 turns from

the analyst running the tests to the reader interpreting them.

3.1 *A Formal Decision Problem*

Consider a researcher who tests m null hypotheses $H_1, \ldots, H_m$ against two-sided alternatives. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \{0,1\}^m$ denote the true state of nature: $\theta_j = 0$ if H_j is true and $\theta_j = 1$ if H_j is false. Based on observed data $\mathbf{X}$, the researcher produces a vector of decisions $\mathbf{d} = (d_1, \ldots, d_m) \in \{0,1\}^m$, where $d_j = 1$ means "reject H_j." A decision-maker uses the rejection vector to choose an action $a(\mathbf{d})$ from some action set $\mathcal{A}$, with payoff

$$U = U\big(a(\mathbf{d}),\, \boldsymbol{\theta}\big). \tag{6}$$

Two types of errors are possible. A *false rejection* occurs when $d_j = 1$ but $\theta_j = 0$. A *missed detection* occurs when $d_j = 0$ but $\theta_j = 1$. The consequences of these errors, and therefore the appropriate error criterion, depend entirely on the mapping from $\mathbf{d}$ to a.

Example 1: Program adoption (confirmatory). A government evaluates a job training program on m labor-market outcomes. The decision is binary: adopt or do not adopt. A simple rule for our purposes is to adopt if at least one test rejects, that is, if the evidence indicates the program affects some outcome. Suppose the program is completely ineffective ($\boldsymbol{\theta} = \mathbf{0}$). A false adoption occurs whenever at least one $d_j = 1$. Let $V = \sum_{j:\,\theta_j = 0} d_j$ denote the number of false rejections. Under this binary action the expected loss is linear in the indicator that any false rejection occurs:

$$L_I \;=\; C_I \cdot \mathbb{P}(V \geq 1), \tag{7}$$

where $C_I > 0$ is the cost of adopting an ineffective program. On the benefit side, a correct adoption yields benefit $B > 0$. The decision-maker's problem is

$$\min_{\mathbf{d}} \;\; C_I \cdot \mathbb{P}(V \geq 1 \mid \boldsymbol{\theta} = \mathbf{0}) \quad \text{subject to} \quad \mathbb{P}\Big(\textstyle\sum_{j=1}^{m} d_j \geq 1 \mid \boldsymbol{\theta} \neq \mathbf{0}\Big) \geq \kappa, \tag{8}$$

where $\kappa \in (0,1)$ is a target power level, defined as the probability of detecting at least one true effect when some hypothesis is false. The quantity $\mathbb{P}(V \geq 1)$ is precisely the *familywise error rate* (FWER). Requiring

$$\text{FWER} \;=\; \mathbb{P}(V \geq 1) \;\leq\; \alpha \tag{9}$$

directly bounds the probability of a wrong aggregate decision. FWER control is the natural consequence of a binary decision problem in which even a single false positive can lead to a wrong action.

Example 2: Screening potential mechanisms (exploratory). A development economist studies a community health intervention and wants to understand *which* of m potential mechanisms (nutrition, sanitation, maternal health, vaccination uptake, school attendance) are actually affected. The goal is not a single binary decision but the identification of a promising set for further investigation. The researcher's output is the set of rejected hypotheses, $\mathcal{R} = \{j : d_j = 1\}$.

A false rejection means one mechanism is mistakenly flagged for follow-up. Suppose each false lead costs $c > 0$ and each true discovery yields benefit $b > 0$. Let V and S denote the number of false and true discoveries, and $R = V + S$ the total number of rejections. The expected net benefit per discovery is

$$\mathbb{E}\left[\frac{b \cdot S - c \cdot V}{R} \,\middle|\, R > 0\right] \;=\; b - (b+c)\,\mathbb{E}\left[\frac{V}{R} \,\middle|\, R > 0\right]. \tag{10}$$

The quality of the screening output depends on the false discovery proportion V/R. Controlling its expectation,

$$\text{FDR} \;=\; \mathbb{E}\left[\frac{V}{\max(R,1)}\right] \;\leq\; q, \tag{11}$$

bounds the expected fraction of false leads. A researcher who controls FDR at $q = 0.10$

is saying: "Of the mechanisms I flag, I expect at most ten percent to be false leads." This is a statement about the average quality of the output, quite different from the FWER guarantee, which is a statement about the probability of *any* error at all.

The error criterion follows from the decision problem. When multiple tests feed into a binary action, even one false positive can lead to a wrong decision, and the problem calls for FWER control. When the tests screen candidates, the relevant quantity is the proportion of false leads, and FDR is the natural object to control.

The framework in (6)–(8) does more than label FWER and FDR as error criteria. Minimizing expected loss also pins down the significance level α itself. Take a single-test version of Example 1 with statistic $t \sim \mathcal{N}(0,1)$ under H_0 and $t \sim \mathcal{N}(\mu,1)$ under H_1, and decision rule "adopt if $t \geq c$." The expected net loss as a function of the rejection threshold is

$$L(c) \;=\; C_I \cdot \big(1 - \Phi(c)\big) \;-\; B \cdot \Phi(\mu - c), \tag{12}$$

where Φ denotes the standard normal CDF. Differentiating and setting to zero gives the first-order condition $C_I \cdot \phi(c) = B \cdot \phi(\mu - c)$, which solves to

$$c^* \;=\; \frac{\mu}{2} + \frac{1}{\mu}\,\ln\!\left(\frac{C_I}{B}\right), \qquad \alpha^* \;=\; 1 - \Phi(c^*). \tag{13}$$

The optimal threshold rises in the cost ratio C_I/B and falls in the signal strength μ. At $C_I = B$ the threshold lies exactly halfway between the null and alternative means ($c^* = \mu/2$), and at $C_I/B = 20$ with $\mu = 3$, it rises to $c^* \approx 2.5$, which corresponds to $\alpha^* \approx 0.006$. The conventional $\alpha = 0.05$ is therefore not a primitive. It corresponds to a particular implicit cost ratio, one that a decision-maker facing a materially different trade-off may reasonably reject.

Under m-test FWER control, the analysis extends with per-test thresholds determined by the joint distribution of the test statistics. With independent tests the FWER as a function of the per-test threshold c is $1 - \Phi(c)^m$ under the global null, and the loss function

generalizes to

$$L_m(c) \;=\; C_I \cdot \left[1 - \Phi(c)^m\right] \;-\; B \cdot \left[1 - (1 - \Phi(\mu - c))^m\right], \tag{14}$$

whose first-order condition $C_I \cdot \Phi(c)^{m-1} \cdot \phi(c) = B \cdot (1 - \Phi(\mu - c))^{m-1} \cdot \phi(\mu - c)$ does not in general admit a closed form. The qualitative conclusion of the single-test case carries through. The optimal threshold rises with m, the per-test α^* falls, and both move with the cost ratio C_I / B and the signal strength μ as before. Under positive dependence the resampling procedures of Section 4 achieve FWER control at less conservative thresholds, so the Bonferroni bound is conservative for applied work. The familiar complaint that multiple testing "sacrifices power" treats α as fixed, but the decision-theoretic framework treats it as chosen. The right response to low power in a setting where false leads are cheap is to raise α, not to abandon the adjustment.

The decision-theoretic framing in (6)–(14) takes the decision rule and loss function as primitives and derives the optimal critical value. Viviano et al. (2026) go a step further upstream: in their principal–agent model a planner designs the testing rule and a researcher chooses how many hypotheses to test, and the optimal correction depends on how the researcher's costs scale with the number of tests. When marginal costs per hypothesis are low, the incentive to expand the testing portfolio makes Bonferroni-type corrections the planner's optimal response; when they are high, the cost structure already disciplines the number of tests and no correction is needed. Their calibrations from clinical-trial and program-evaluation data place real-world settings between these extremes. The two foundations are complementary—ours derives the optimal threshold conditional on a fixed family, theirs determines when and how much correction is warranted in the first place—and together they imply that conventional thresholds encode a particular implicit position on both the loss function and the research cost structure, from which researchers in materially different environments may reasonably depart.

Empirical papers often carry both confirmatory and exploratory questions at once. A randomized evaluation might pre-specify two primary outcomes while also examining ten secondary outcomes and several subgroups, in which case the primary family calls for FWER and the exploratory family calls for FDR. Hierarchical testing procedures (Section 4.4) operationalize this broader logic of allocating different error budgets to different layers, concentrating the budget on primary claims.

3.2 *Error Criteria*

The researcher tests m null hypotheses, of which $m_0 \leq m$ are true. Let V denote the number of false rejections, S the number of correct rejections, and $R = V + S$. Three standard criteria are:

FWER: $\mathbb{P}(V \geq 1) \leq \alpha$ — probability of *any* false rejection. Natural for confirmatory binary decisions.

FDR: $\mathbb{E}\left[V / \max(R, 1)\right] \leq q$ — expected share of false discoveries among rejections. Natural for screening problems.

k-FWER: $\mathbb{P}(V \geq k) \leq \alpha$ — probability of k or more false rejections. An intermediate criterion that tolerates a small known number of false positives.

A procedure provides *strong* FWER control if $\mathbb{P}(V \geq 1) \leq \alpha$ for every configuration of true and false nulls, and *weak* control if this holds only under the global null. Strong control is the standard for applied work, where the researcher cares which specific hypotheses are rejected, not merely whether any false rejection occurred. Under the global null, controlling FDR at level q implies FWER $\leq q$, so the two coincide. Away from the global null FDR is a less stringent criterion, appropriate for screening.

A complementary concept that runs through the rest of the paper is the *adjusted p-value*. For any given multiple testing procedure (MTP), the adjusted p-value for hypothesis H_j is

the smallest significance level at which the procedure would reject H_j:

$$\tilde{p}_j \;=\; \inf\{\alpha \in [0,1] : H_j \text{ is rejected by the MTP at level } \alpha\}. \tag{15}$$

Adjusted p-values preserve the familiar interpretation: reject H_j at level α if and only if $\tilde{p}_j \leq \alpha$. They also facilitate comparison across procedures. A Romano–Wolf adjusted p-value of 0.03 for H_j means that even after accounting for all m tests, the evidence against H_j is sufficient to reject at the three percent level.

3.3 *Interpreting Individual Hypotheses*

A question arises naturally from the preceding discussion but has received little formal attention in applied economics: how should a reader who cares about a *specific* hypothesis H_j interpret the results when m tests have been conducted? Suppose a labor economist reads a paper evaluating a job training program on $m = 8$ outcomes and is specifically interested in the effect on earnings (H_1). The unadjusted p-value for H_1 is 0.03. Should she take this at face value, or should the fact that seven other hypotheses were tested raise her posterior on H_1, lower it, or leave it unchanged?

The answer depends on the reader's prior beliefs about the joint distribution of *truth states* across hypotheses, an object distinct from the dependence among test statistics. Two extremes bracket the question. If the reader believes truth states are independent across hypotheses, the other tests are uninformative about H_1 and multiplicity is irrelevant from her perspective. If she believes they are perfectly correlated, the individual test of H_1 is redundant. Both extremes are unrealistic. Independence with prior probability $1/2$ on each hypothesis implies a prior on the global null of just $(1/2)^8 \approx 0.4\%$, while perfect correlation rules out the plausible scenario in which a program affects earnings but not health. An intermediate prior is needed, but specifying one requires substantive beliefs about how the program operates, and those beliefs are hard to formalize and arbitrary for any particular choice.

A minimal formal model clarifies what is identifiable. Suppose the truth states $\theta_1, \ldots, \theta_m$ are exchangeable conditional on an unknown common proportion π of true alternatives, with $\theta_j \mid \pi \overset{iid}{\sim} \text{Bernoulli}(\pi)$. The exchangeable structure isolates the role of m on identifiability and is not meant as a descriptive claim. Conditional on $\theta_j = 0$ the p-value p_j is uniform on $[0, 1]$, and conditional on $\theta_j = 1$ it follows some alternative density f_1. By Bayes' rule,

$$\mathbb{P}(\theta_j = 1 \mid p_1, \ldots, p_m) \;=\; \int \frac{\pi\, f_1(p_j)}{\pi\, f_1(p_j) + (1 - \pi)}\, d\mathbb{P}(\pi \mid p_1, \ldots, p_m). \tag{16}$$

The other p-values enter only through the posterior on π: when m is large the data pin down π sharply and the posterior on θ_j depends primarily on p_j itself, which is the regime in which the *local false discovery rate* of Efron (2010) becomes computable and provides a Bayesian answer to the question. When m is small, as in the typical economics regime, the data do not pin π down and the reader's own prior continues to matter.

To fix ideas, return to the labor economist with $m = 8$ and $p_1 = 0.03$. Take the alternative density induced by a moderate effect: under $\theta_j = 1$ the test statistic is $z_j \sim \mathcal{N}(2, 1)$, which delivers about 64% power at $\alpha = 0.05$. Then $f_1(0.03) \approx 5.8$, and with prior $\pi = 0.3$ on the proportion of true alternatives the conditional posterior is

$$\mathbb{P}(\theta_1 = 1 \mid p_1, \pi = 0.3) \;=\; \frac{0.3 \cdot 5.8}{0.3 \cdot 5.8 + 0.7} \;\approx\; 0.71.$$

A reader with a more skeptical prior $\pi = 0.1$ obtains 0.39; a reader with $\pi = 0.5$ obtains 0.85. With $m = 8$ the other seven p-values barely move the posterior on π, so the spread across prior choices dominates the posterior.

The practical recommendation has two parts. For authors: report unadjusted p-values, adjusted p-values, and the family over which the adjustment was carried out, and prespecify that family before the data are seen (Section 5). For readers: the adjusted p-value is the right benchmark when no outcome-specific prior is in play — the default in

confirmatory work, where readers do not bring a model that distinguishes one outcome from another. When the reader does bring such a prior, equation (16) is the object to update, and the calculation above shows that in the small-m regime that prior is doing most of the work. For families large enough that the alternative density can be estimated from the p-value distribution itself, the local-FDR machinery of Efron (2010) and the asset-pricing application of Harvey et al. (2026) provide a Bayesian benchmark that bypasses the need for a subjective prior.

4 Methods and Recommendations

Section 3 fixes the error criterion the procedure must deliver. The procedure that delivers it depends on two further inputs: the dependence structure among test statistics, which determines whether an analytical bound or a resampling estimator is appropriate, and any prior ordering on the hypotheses, which determines whether the error budget should spread evenly or concentrate. The remaining subsections develop procedures along these lines—analytical adjustments first, then resampling methods that exploit dependence, then hierarchical constructions that exploit ordering, then simultaneous confidence intervals for magnitudes. Lastly, we consolidate the recommendations.

4.1 *Bonferroni and Stepwise Adjustments*

The simplest analytical adjustment is the Bonferroni correction, which tests each of m hypotheses at significance level α/m. It controls FWER at level α under any dependence structure, by the union bound:

$$\mathbb{P}(V \geq 1) \;\leq\; \sum_{j:\,\theta_j=0} \mathbb{P}(p_j \leq \alpha/m) \;\leq\; m_0 \cdot (\alpha/m) \;\leq\; \alpha.$$

This generality is the source of both its appeal and its limitation. The union bound is tight only when test statistics are close to independent. When they are positively correlated, Bonferroni gives up power without compensating gains. The Šidák correction replaces

α/m with $1-(1-\alpha)^{1/m}$ and is slightly less conservative, but it requires independence for exact control, and the difference between the two is negligible next to the conservatism they share.

Holm's stepdown procedure (Holm, 1979) orders the m p-values from smallest to largest, $p_{(1)} \leq \cdots \leq p_{(m)}$, and compares the i-th smallest to $\alpha/(m-i+1)$, stopping at the first non-rejection. The adjusted p-values are

$$\tilde{p}_{(i)}^{\text{Holm}} \;=\; \min\Big(\max_{k\leq i}\big\{(m-k+1)\,p_{(k)}\big\},\;1\Big). \tag{17}$$

Holm controls strong FWER under *any* dependence structure and *uniformly dominates* Bonferroni: it rejects everything Bonferroni rejects and sometimes more. The computational cost is negligible. In practice we see little reason to use raw Bonferroni or Šidák when Holm is available, unless the objective calls for a single-threshold rule that can be stated with one number.

Hochberg's step-up procedure (Hochberg, 1988) reverses the direction and is more powerful than Holm when valid. Its validity rests on the Simes inequality on ordered p-values, which can fail under negative or complex dependence, and the failure is hard to diagnose ex ante in confirmatory FWER settings. Holm is therefore the safer default; Hochberg is a defensible choice only when the dependence can be shown to be positive and simple enough for Simes to hold. For the related criterion of k-FWER, introduced in Section 3.2, Lehmann and Romano (2005) and Romano and Wolf (2010) generalize Holm.

Holm's limitation is that it ignores the dependence structure among test statistics; when outcomes are positively correlated, that costs power. The resampling methods of Section 4.3 exploit the dependence directly.

4.2 *False Discovery Rate Methods*

When the goal is exploratory, controlling the expected proportion of false discoveries fits the problem better than controlling the probability of any false discovery at all. The proce-

dure that delivers this control is the Benjamini–Hochberg rule (Benjamini and Hochberg, 1995): it orders the p-values and finds the largest k such that

$$p_{(k)} \;\le\; \frac{k}{m}\, q, \tag{18}$$

where q is the target FDR level, then rejects $H_{(1)}, \ldots, H_{(k)}$. The adjusted p-values are

$$\tilde{p}^{\mathrm{BH}}_{(i)} \;=\; \min\Bigl(\min_{k \ge i}\Bigl\{\frac{m}{k}\, p_{(k)}\Bigr\},\; 1\Bigr). \tag{19}$$

BH controls FDR under independence and under *positive regression dependence on a subset* (PRDS, Benjamini and Yekutieli, 2001), a technical condition satisfied by multivariate normal test statistics with nonnegative correlations. Since positive dependence is the typical case in applied work, BH is valid in most applications, with the qualification that PRDS may fail in settings with negative dependence or complex error structures.

The Benjamini–Yekutieli (BY) procedure replaces the BH threshold with a more conservative one that divides by the harmonic number $\mathcal{H}_m = \sum_{i=1}^{m} 1/i$, ensuring FDR control under arbitrary dependence. For $m = 20$, the harmonic number is approximately 3.60, so BY shrinks every threshold by more than a factor of three. This conservatism is the price of validity without dependence assumptions.

Adaptive methods that estimate the proportion of true nulls $\pi_0 = m_0/m$ can sharpen the BH thresholds (Storey, 2002) but need m large enough for reliable estimation, which limits their use in economics. A useful by-product is the *q-value* (Storey, 2002)—the smallest level at which a hypothesis would be declared significant—which, reported alongside the rejection set, lets readers gauge how stable the conclusions are to the chosen tolerance.

Two cautions: FDR is an expectation over hypothetical replications, not a guarantee for any given study—and when rejections are few the realized false-discovery proportion is a coarse, discrete random variable—and FDR does not bound the probability of *any* false rejection, which is why confirmatory settings call for FWER control. With those

qualifications in mind, BH is the default for exploratory analyses: simple, powerful, and valid under positive dependence.

4.3 *Resampling-Based Methods*

Instead of using analytical bounds that assume worst-case dependence, resampling methods *estimate* the joint distribution of the test statistics under the null hypothesis from the data themselves. The first procedure to implement this idea was the Westfall–Young (WY) stepdown (Westfall and Young, 1993), which generates B resampled datasets under the null, via permutation or a constrained bootstrap, and computes the resampled p-values $p_1^{*(b)}, \ldots, p_m^{*(b)}$ in each. For every hypothesis $H_{(i)}$ in the ordered sequence, it then forms the successive minimum $q_{(i)}^{*(b)} = \min_{k \geq i} p_{(k)}^{*(b)}$ and the raw adjusted p-value

$$\tilde{p}_{(i)}^{\mathrm{WY,raw}} \;=\; \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{ q_{(i)}^{*(b)} \leq p_{(i)} \}, \tag{20}$$

imposes monotonicity through $\tilde{p}_{(i)}^{\mathrm{WY}} \leftarrow \max\big(\tilde{p}_{(i-1)}^{\mathrm{WY}}, \tilde{p}_{(i)}^{\mathrm{WY,raw}}\big)$, and rejects $H_{(i)}$ whenever $\tilde{p}_{(i)}^{\mathrm{WY}} \leq \alpha$. For the most significant hypothesis the successive minimum runs over all m resampled p-values—the resampled analogue of the maximum statistic under the null—and over a shrinking set thereafter, reflecting the reduced burden after earlier rejections. Westfall–Young provides strong FWER control under a technical condition called *subset pivotality*: the joint distribution of the test statistics for a subset of true nulls must not depend on which of the remaining hypotheses are true or false. The condition holds automatically when the test statistics are independent, or when studentization (dividing each estimate by its own standard error) uses equation-by-equation variance estimates. It can fail in common applied settings.[2]

[2]Consider a seemingly unrelated regression system $y_j = \mu_j + \beta_j T + \varepsilon_j$ studentized with a *pooled* variance estimator $\hat{\sigma}^2 = \frac{1}{nm} \sum_{i,j} \hat{\varepsilon}_{ij}^2$ and resampled under the joint null. An equation with $\beta_k \neq 0$ leaves a $\beta_k T_i$ component in the residuals, inflating $\hat{\sigma}^2$, so the studentized statistic for a true null $\beta_j = 0$ has a null distribution that shifts with the non-null parameters elsewhere—exactly the truth-status dependence subset pivotality rules out. Equation-by-equation variance estimators restore pivotality at the cost of efficiency.

The Romano–Wolf procedure (Romano and Wolf, 2005a,b, 2016) relaxes the pivotality requirement in two ways. First, the null distribution is simulated using a bootstrap that *centers at the observed estimates* and uses studentized test statistics, so the bootstrap distribution reflects the actual variance structure of the data rather than imposing all nulls simultaneously. Second, after rejecting a hypothesis, Romano–Wolf computes the next critical value from the joint distribution of the remaining test statistics, further improving power by reducing the effective number of comparisons at each step. The procedure provides strong FWER control asymptotically under regularity conditions satisfied in most applied settings and handles heteroskedasticity, clustering, and other features of economic data naturally through the bootstrap.

In experimental economics, List et al. (2019) brought this resampling approach into standard practice: their procedure controls the familywise error rate across the multiple outcomes, treatment arms, and subgroups a single experiment generates by bootstrapping their joint distribution, and the accompanying `mhtexp` command made it routine. The default we advocate is the same in spirit; what the framework of this paper adds is the step before the procedure—deciding when familywise control is the right target at all—and the extension beyond the experimental case to the other designs of Section 2, to hierarchical orderings, and to magnitudes.

The power advantage runs through the distribution of the maximum statistic: the FWER critical value is the $(1-\alpha)$-quantile of $\max_j |t_j|$ under the null. Bonferroni and Holm bound it under worst-case dependence and so are too conservative when the actual dependence is positive; Romano–Wolf bootstraps it directly, so the two coincide at $\rho = 0$ and diverge as positive dependence grows.

4.4 *Hierarchical Methods*

A hierarchy is present whenever the researcher can credibly distinguish primary hypotheses from secondary (mechanism or mediator) and exploratory hypotheses. The Piso Firme evaluation of Cattaneo et al. (2009) provides one example: the program subsidized cement

floors, so the natural ordering tests floor coverage first (the mechanism), then child health (the target), then household wellbeing (downstream)—a parallel-gatekeeping ordering across the three families. Treating all hypotheses as a flat family applies the same multiplicity penalty to primary and exploratory tests alike, giving up power that hierarchical procedures could direct toward primary claims.

The simplest hierarchical approach is the fixed-sequence procedure: test $H_1, \ldots, H_m$ in a pre-specified order, each at the full level α, stopping at the first non-rejection—the primary hypothesis faces no multiplicity penalty, but a non-rejection blocks everything after it. Fallback procedures (Wiens, 2003) relax this by assigning initial weights and redistributing a rejected hypothesis's significance to those remaining (α-propagation), so testing can continue past a non-rejection. Gatekeeping procedures lift the idea from individual hypotheses to families: family $\mathcal{F}_{k+1}$ is tested only if a gate condition holds in $\mathcal{F}_k$—*serial* gatekeeping requires all of $\mathcal{F}_k$ to reject, *parallel* gatekeeping only one. Bretz et al. (2009) unify fixed-sequence, fallback, and gatekeeping procedures as special cases of a graphical framework that represents them as directed weighted graphs over hypotheses.

The researcher can use Romano–Wolf within each family and a gatekeeping structure across families. The combination inherits the strengths of both approaches. The hierarchy concentrates the error budget on primary hypotheses, and the resampling step exploits dependence within each family. This is our recommended strategy when hypotheses come with a natural hierarchy. Its power depends on the hierarchy being correctly specified. When the hierarchy matches the underlying causal structure, the gains can be substantial. When it does not—for instance, when the true effects are concentrated in the exploratory family rather than the primary family—the primary gate may fail to open. The hierarchical procedure can then end up with near-zero power where the effects are real, while flat Romano–Wolf would detect them. The hierarchy has to be set by substantive reasoning about the causal structure, not by looking at the results.

A complementary structural choice within a hierarchy is to test a summary index as

the primary hypothesis for a family of related outcomes (Kling et al., 2007; Anderson, 2008; Anderson and Magruder, 2022), with the within-domain outcomes tested only if the index rejects. This is one way of building a within-family hierarchy; the illustration of Section 6.4 instead treats each family as a flat set of outcomes.

The construction provides strong FWER control by the standard closed-testing argument, provided that the within-family Romano–Wolf tests control FWER at level α and the hierarchy satisfies a monotone-effects condition: a true non-null at any level implies a true non-null at every higher level.

4.5 *Simultaneous Confidence Intervals*

Multiple-testing adjustment, as developed in Sections 4.1–4.4, controls the probability of false rejections across a family of hypotheses. A program's policy relevance depends on how large its effects are, not only on whether they are statistically distinguishable from zero, and a serious report should communicate both. Simultaneous confidence intervals are the magnitude analogue of multiple-testing-adjusted tests: a set of m intervals $\{[\hat{\beta}_j \pm c_j]\}_{j=1}^{m}$ such that all m contain their target parameters jointly with probability at least $1-\alpha$. Reporting m individual 95% intervals does not deliver this guarantee—the joint coverage rate falls with m under positive dependence in exactly the way the unadjusted FWER rises. Constructing a single critical value that delivers joint coverage is the same problem as constructing an FWER-valid test, and the same dependence-aware machinery solves both.

The maxT bootstrap delivers a sharp critical value c via the same resamples used for Romano–Wolf testing: generate B bootstrap samples, compute the studentized maximum $T_b^* = \max_j |\hat{\beta}_j^{*b} - \hat{\beta}_j| / \mathrm{SE}(\hat{\beta}_j^{*b})$ in each, take c as the $(1-\alpha)$-quantile of $\{T_1^*, \ldots, T_B^*\}$, and report $[\hat{\beta}_j \pm c \cdot \mathrm{SE}(\hat{\beta}_j)]$. Because the bootstrap captures the dependence structure, these intervals are narrower than Bonferroni intervals (each constructed at the $1-\alpha/m$ level) under positive correlation, a gain quantified in the simulations of Section 6. For the single-step maxT with a common critical value c, there is an exact duality with adjusted testing—a

simultaneous interval for β_j excludes zero if and only if the corresponding test rejects H_j at level α—while the stepdown version uses step-specific critical values, so the duality is more intricate.

4.6 *Practical Guidance*

Three method-application errors recur in applied work, each from a choice made by habit rather than derived from the decision the tests inform. **Using raw Bonferroni when Holm is available** wastes power for nothing, since Holm dominates it at no cost. **Confusing FDR with FWER** applies the wrong guarantee—BH to a confirmatory endpoint, or Bonferroni to a large exploratory screen; the criterion should follow the decision. **Pairing an FWER-adjusted test with unadjusted intervals** understates the joint uncertainty the correction was meant to address; magnitudes call for simultaneous intervals at the test's level. Making these choices deliberately is a matter of a short sequence of decisions:

1. *Do the tests inform a common conclusion?* If not, single-test inference applies; if so, proceed.
2. *What is the family?* Every hypothesis that collectively informs the conclusion, fixed in the pre-analysis plan.
3. *What is the criterion?* Binary decisions $\Rightarrow$ FWER; screening $\Rightarrow$ FDR; mixed $\Rightarrow$ FWER on primary, FDR on exploratory.
4. *Which procedure?*
    - For FWER: Romano–Wolf by default; hierarchical Romano–Wolf (Section 4.4) when an ordering is credible; the wild cluster bootstrap of Cameron et al. (2008) when clusters are few ($G \lesssim 30$); Holm as fallback in very small samples ($n < 50$) or when studentization is not asymptotically pivotal.
    - For FDR: Benjamini–Hochberg at $q = 0.05$ or $q = 0.10$; reserve Benjamini–Yekutieli for adversarial dependence.

When Romano–Wolf is used for testing, the natural CI partner is the maxT bootstrap at the same nominal level; with analytical methods, Bonferroni intervals at $1 - \alpha/m$ are the default companion.

Resampling should use at least $B = 1{,}000$ draws, with the adjusted p-values verified to be monotone. The report presents both unadjusted and adjusted p-values, the method, the family, any hierarchical structure, and simultaneous confidence intervals when feasible. Mature software in Stata, R, and Python (Table 1, Panel A) implements the full menu; Romano–Wolf at this scale runs in seconds. The Stata command `mht` (Calonico et al., 2026) packages the full Section 4 menu in one interface, with hierarchical gatekeeping via the `mht hier` subcommand and a cross-platform C plugin for the wild cluster bootstrap.[3] It is the implementation used in the empirical illustration of Section 6.4.

Pre-specification discipline—selective reporting, post-hoc method choice, and undeclared non-adjustment—is addressed separately in Section 5.

5 Pre-specification and Multiple Testing

The methods reviewed in Section 4 control error rates given a fixed family of hypotheses. The researcher, however, chooses which outcomes to test, how to group them, which error criterion to apply, and which adjustment method to use. Each of these choices affects the final results, and each can be made before or after examining the data. Pre-specification of the testing plan is the necessary complement to multiple testing adjustment in confirmatory work. Exploratory analyses, by design, cannot be fully pre-specified; they should be labeled as such, and confirmatory inference reserved for follow-up data. The rest of this section sets out the specific vulnerabilities that arise without pre-specification, and locates within-study multiplicity in the broader context of the replication crisis (Christensen and Miguel, 2018).

A multiple testing procedure adjusts only for the hypotheses it is told about. If the

[3]The command and the replication code are available at `https://www.sebastiangaliani.com/academic-code`.

family is incomplete, or is selected to make the results look favorable, the error guarantee is void. Pre-analysis plans close this gap by fixing the family, the error criterion, the adjustment method, and any hierarchical ordering before the data are examined. The case for pre-specification does not rest on suspicion of misconduct. Even a researcher acting in good faith faces a stream of decisions during analysis. Whether to include a borderline outcome, how to handle an unplanned subgroup, whether to split a large family: any of these can distort the error guarantees without anyone intending it—the "garden of forking paths," in which the realized analysis is contingent on the data even absent any deliberate search (Gelman and Loken, 2013). Pre-specification removes those decisions from the post-data environment, where they are most exposed to influence by what the data already show. Choosing the adjustment method after seeing the p-values is a form of p-hacking that voids the error guarantee, regardless of whether the chosen procedure is itself valid.

A formal version of the same point appears in Viviano et al. (2026): in their principal–agent model, the optimal testing rule depends on the researcher's cost structure and is implementable only when the family and the rule are committed to in advance. Pre-specification is the institutional implementation of that commitment, valuable even under the presumption of good faith because cost-structure incentives operate independently of intent.

Each class of procedure has its own failure mode when the testing plan is not fixed in advance. FWER procedures are most directly compromised by selective reporting (testing m hypotheses and reporting only the m' with the smallest p-values). FDR procedures are compromised both by omitting tested hypotheses and by *adding* easy-to-reject true alternatives that inflate the denominator R, diluting V/R in the reported set and leaving room for a high V without breaching the nominal threshold. Hierarchical procedures fail when the ordering is chosen ex post to place the strongest findings at the top, which voids the hierarchical error guarantee and is why Section 4.4 insists on substantive justification for any hierarchy. In every case the remedy is the same: pre-specify the family, the criterion,

the method, and any ordering. Report any deviation from the plan with its rationale.

Selection makes the realized error rate diverge sharply from the nominal. A researcher who tests $m = 20$ hypotheses under the global null, keeps only the $m' = 5$ with the smallest p-values, and applies Bonferroni at level α/m' to the reported subset claims FWER control at $\alpha = 0.05$. Under independence the true FWER is $1 - (1 - 0.05/5)^{20} \approx 0.18$, more than three times nominal. The distortion grows with the ratio m/m' and shrinks but does not vanish under positive dependence. Error control depends on the number of tests *conducted*, not the number reported. Pre-specification closes the gap by fixing m in advance, and registered reports close it further by making the conducted set verifiable.

In practice, pre-specification operates through pre-analysis plans (PAPs; Olken, 2015) registered with bodies such as the AEA RCT Registry, AsPredicted, or the Open Science Framework. A useful PAP also commits to a target level on the chosen error criterion and to any hierarchical ordering across families. Deviations from a registered plan are not failures but part of how research adapts to surprises in the data; they should be reported transparently, with pre-specified and post-hoc results presented separately. The same transparency applies to the decision not to adjust: declining a correction can be defensible—a small pre-specified primary endpoint, or tests of genuinely separate questions—but the choice should be explicit and reasoned, since a tacit acceptance of an aggregate error rate of $1 - (1 - \alpha)^m$ is still a decision.

The multiplicity discussed in this paper is within-study: it arises when a single research design generates many tests that jointly inform one decision. A parallel across-study problem arises because the publication process selects on significance, so that even when every individual paper controls its within-study error rate, the published literature can contain far more false positives than α (Simonsohn et al., 2014; Andrews and Kasy, 2019; Brodeur et al., 2020).

Journals and referees prefer findings that are statistically significant, so the distribution of published estimates is the conditional distribution given that the test rejected, not the

unconditional sampling distribution; the conditional mean is biased away from zero, and the share of false positives among published findings can substantially exceed α. Brodeur et al. (2020) document the resulting excess mass of p-values just below conventional thresholds across the published economics literature, and Andrews and Kasy (2019) provide an estimator that corrects published effects for the implied selection. Within-study multiple testing adjustment cannot remove the journal-stage filter, but it changes what the filter operates on. Dependence-aware procedures such as Romano–Wolf lower the within-study rejection rate under the null, so the supply of false positives entering the publication process is smaller and the published distribution is correspondingly less distorted. Pre-specification carries the commitment device upstream: by fixing the family and the method before the data are seen, it removes the within-paper analogue of publication bias — selective reporting — and registered reports extend that commitment to the journal level. Tackling within-study multiplicity rigorously is an upstream complement to the across-study corrections of Andrews and Kasy (2019) and Brodeur et al. (2020), not a substitute for them.

6 NUMERICAL ILLUSTRATIONS

This section provides simulation evidence for the recommendations of Section 4. We compare the principal procedures under two flat-dependence designs — AR(1) and equicorrelated (Table 2) — and a three-level hierarchical design (Table 3), with $n = 400$ observations and $m = 20$ hypotheses throughout. Section 6.3 then stress-tests the flat designs against clustering, heavy tails, and mixed-sign dependence, and Section 6.4 carries the full framework through the Piso Firme evaluation of Cattaneo et al. (2009) (Table 4).

6.1 *Simulation Design*

We consider a standard multi-outcome regression setup with $n = 400$ observations and $m = 20$ null hypotheses $H_j : \beta_j = 0$, where β_j is the coefficient on a binary treatment

indicator:

$$y_{ij} \;=\; \mu_j + \beta_j T_i + \varepsilon_{ij}, \qquad j = 1, \ldots, m. \tag{21}$$

Treatment is random: $T_i \sim \text{Bernoulli}(0.5)$. The error vector follows a multivariate normal with one of two covariance structures. The AR(1) design uses $\Sigma_{jk} = \rho^{|j-k|}$, generating positive dependence that decays with distance in the ordering (neighbors strongly correlated, distant outcomes near-independent). The equicorrelated design uses $\Sigma_{jk} = \rho$ for $j \neq k$ and $\Sigma_{jj} = 1$, giving uniform pairwise correlation—a single-factor structure in which every outcome loads on a common factor. The two bracket the dependence patterns common in applied work: AR(1) for outcomes ordered along a time or spatial axis, equicorrelated for the "multiple outcomes, same treatment" case in which related measures share a latent effect.

We consider $\rho \in \{0, 0.3, 0.7\}$ for each structure and three signal configurations: *global null* ($\beta_j = 0$ for all j), *sparse signal* (20% of coefficients non-null at $\delta = 0.25$), and *dense signal* (50%). Non-null positions are drawn at random and held fixed across structures and correlation levels, so that variation in the output reflects the dependence alone. For each configuration we run 1,000 Monte Carlo replications of $m = 20$ OLS regressions and apply unadjusted testing, Bonferroni, Šidák, Holm, Hochberg, Romano–Wolf, Westfall–Young, Benjamini–Hochberg, and Benjamini–Yekutieli at $\alpha = 0.05$ (Romano–Wolf and Westfall–Young with $B = 1{,}000$ resamples, seeded for reproducibility), reporting FWER, power, and FDR.

To illustrate the value of hierarchical testing (Section 4.4), we also simulate a three-level hierarchical design with the same $m = 20$ hypotheses. The DGP is unchanged, but the hypotheses are grouped to mirror a primary–mechanism–exploratory structure. Level 1 is a primary family of two outcomes with strong effects ($\delta_1 = 0.40$). Level 2 is a mechanism family of eight outcomes: four receive moderate effects ($\delta_2 = 0.25$) and four are null. Level 3 is an exploratory family of ten outcomes: three receive weak downstream effects ($\delta_3 = 0.15$) and seven are null. Nine hypotheses are non-null in total, and the

hierarchy requires only $\delta_1 > \delta_2 > \delta_3 > 0$. We fix the correlation among errors at $\rho = 0.5$ and compare flat procedures (unadjusted, Bonferroni, Šidák, Holm, Hochberg, Romano–Wolf, Westfall–Young, Benjamini–Hochberg, Benjamini–Yekutieli) with the hierarchical gatekeeping counterparts of Holm and Romano–Wolf. We report the global FWER, level-specific power (the rejection rate averaged over the non-null hypotheses within each level), and overall power (the rejection rate averaged over all nine non-null hypotheses).

6.2 *Results*

The simulations bear out the framework's three claims: every adjusted procedure controls its stated criterion, resampling converts positive dependence into power gains over the analytical methods, and hierarchical gatekeeping recovers further power at the primary and mechanism layers. Table 2 reports the two flat designs.

Across the signal panels of Table 2, every FWER procedure holds size at or below the nominal 5% level under both dependence structures, with Benjamini–Yekutieli conservative throughout. The same holds under the global null (omitted from the table for space): adjusted FWER near 0.05 and BY conservative, while the unadjusted procedure shows the inflation predicted by (1)—FWER 0.605 at $\rho = 0$, close to $1 - 0.95^{20}$, declining to 0.437 under AR(1) and 0.265 under equicorrelated dependence at $\rho = 0.7$ as positive dependence shrinks the effective number of independent tests.

The power comparisons confirm the central claim, and reveal that how much resampling helps depends on the shape of the dependence. At independence RW and Holm are indistinguishable (0.306 versus 0.306 sparse, 0.315 versus 0.316 dense), since the bootstrap and the analytical bounds coincide when test statistics are uncorrelated. Under AR(1), which concentrates correlation in neighboring tests, RW's advantage as correlation rises is modest: at $\rho = 0.7$ it reaches 0.336 against Holm's 0.314 under the sparse signal, a 7% gain. Under equicorrelated dependence, where every outcome loads on a common factor—the "multiple outcomes, same treatment" case of Section 2.1—the gain is far larger: 0.393 against 0.293, a 34% gain, with a comparable widening under the dense signal. The

mechanism is the one identified in Section 4.3: RW bootstraps the joint distribution of the maximum statistic directly, so its critical values track the actual dependence, while Bonferroni and Holm rely on bounds calibrated to the worst case. The practical lesson is that resampling pays off most precisely when outcomes share a common driver—the typical multi-outcome setting in applied work.

Benjamini–Hochberg delivers higher power than every FWER procedure—roughly 40% of true alternatives under the AR(1) sparse signal and 52% under the dense one, against RW's 31–34%—but at an FWER of 0.08–0.13, far above nominal. That is no defect: BH controls the FDR, and the FDR column confirms it, holding the false discovery proportion to 0.022–0.039. Under equicorrelated dependence the gap closes as RW exploits the common factor (RW 0.393 versus BH 0.390 at $\rho = 0.7$, sparse). The choice between them is substantive, not methodological—RW when any false positive can mislead a decision, BH when false leads are tolerable in proportion—but either way the power cost of adjustment is real and bounded.

When the study's logic orders hypotheses into primary, mechanism, and exploratory layers, hierarchical procedures concentrate the error budget where it most affects conclusions. Table 3 reports the three-level design at $\rho = 0.5$; all procedures, including the two hierarchical variants, control global FWER near nominal. Flat methods detect the strong Level 1 signals reliably (flat RW power 0.841) but fade at Level 2 (0.313) and collapse at Level 3 (0.077), because the penalty applies uniformly across all twenty hypotheses. Hierarchical gatekeeping opens the primary gate with high probability (L1 power 0.974) and reallocates the budget downstream—a 34% gain at Level 2 and 22% overall (0.430 vs 0.352). Level 3 power is essentially unchanged, since the within-family efficiency gain is offset by the chance that the Level 2 gate fails to open, so hierarchy pays off most at the primary and mechanism layers.

The metrics reported so far—familywise error, power, and the false discovery proportion—all concern hypothesis testing. In most applications the magnitude of an

effect matters as much as whether it is distinguishable from zero, and the same resampling distribution that calibrates the tests also delivers the simultaneous confidence intervals of Section 4.5: in the equicorrelated design at $\rho = 0.7$, the maxT bands are roughly 8% narrower than the corresponding Bonferroni intervals, because the bootstrap exploits the cross-outcome dependence that the analytical construction must ignore. Multiple inference for magnitudes is nonetheless far less developed than for testing—the literature, and this paper, are testing-centric largely by inheritance—and a fuller treatment of simultaneous interval estimation under dependence is an important direction for future work.

6.3 *Robustness*

Three departures from the baseline design leave the conclusions intact (the full panel is reported in the replication package). Under clustered errors—$G = 50$ clusters with intracluster correlation 0.5—a cluster bootstrap delivers nominal FWER for Romano–Wolf, validating the cluster-bootstrap implementation used in Section 6.4, with power comparable to Holm because the cross-outcome dependence runs through the cluster-level shock. Under heavy-tailed $t(5)$ shocks RW keeps its modest advantage, since the studentized statistic remains asymptotically pivotal. And under mixed-sign dependence—a single-factor model with loadings $\lambda_j = \pm\sqrt{0.5}$, so outcomes correlate $+0.5$ within blocks and -0.5 across—RW's edge is largest, 0.334 against Holm's 0.297 under the sparse signal, because the bootstrap captures the cancellation between positively and negatively correlated outcomes that analytical bounds cannot exploit. In every case the unadjusted procedure over-rejects and Benjamini–Yekutieli stays conservative, as in the flat designs.

6.4 *Empirical Illustration: Piso Firme*

We replicate Cattaneo et al. (2009), whose evaluation of the Piso Firme floor-replacement program is the running example of Sections 2.1(a) and 4.4, and work through the multiple-testing workflow on their full set of 17 outcomes. Piso Firme is a binary policy question—the program is either effective enough to scale or not—so the familywise error rate is

the natural criterion. We use Cattaneo et al.'s Model 3 specification with cluster-robust standard errors at the geographic-cluster level (controls listed in the Table 4 notes).

Cattaneo et al. pre-specify three families: HH_floor (cement-floor coverage, the mechanism, $m = 5$), CH_health (child-health outcomes, the target, $m = 7$), and HH_satis (subjective wellbeing, downstream, $m = 5$). We read the evidence two ways. Treated *per family*, each of the three families is adjusted on its own, so the multiplicity argument lives at the family level—where the original study pre-specified it. Treated *hierarchically* (Section 4.4), the causal chain HH_floor $\rightarrow$ CH_health $\rightarrow$ HH_satis orders the families, and a family is read only if the one before it rejects at least one outcome. Within each family we report Holm, Benjamini–Hochberg, and Romano–Wolf with a $B = 1{,}000$ cluster bootstrap (Table 4); implementation details are documented in Calonico et al. (2026).

HH_floor and HH_satis are decisive: every outcome rejects under every adjustment, with Romano–Wolf p-values at the bootstrap floor. Because each of the three families rejects at least one outcome, all three gates open, and the hierarchical reading coincides with the per-family verdict—the causal chain is satisfied, licensing inference from mechanism to target to downstream wellbeing without overturning any family's conclusion. The substantive action sits entirely within CH_health.

Two of the seven child-health outcomes survive within-family adjustment under every method: anemia ($\tilde{p}_{\mathrm{RW}} = 0.037$) and the MacArthur communicative-development score (MCCDTS; $\tilde{p}_{\mathrm{RW}} = 0.020$). Parasitic infestation and the Peabody vocabulary score are significant unadjusted ($p = 0.046$ and 0.031) but do not survive—their Romano–Wolf p-values are 0.236 and 0.206—and diarrhea sits at $p = 0.054$, already above 0.05 before any correction. The two anthropometric z-scores are null at every threshold. The pattern makes the framework's central caution concrete: borderline unadjusted significance is a brittle basis for a claim once the family is tested as a whole.

The final column carries the magnitude reading. The maxT 95% bands, built from the same bootstrap, exclude zero for exactly the twelve outcomes Romano–Wolf rejects (the

duality of Section 4.5), and they report what a test alone cannot: the cement-floor and wellbeing effects are pinned down tightly, while the two surviving child-health effects are bounded only loosely—the MacArthur band runs from 0.67 to 10.44 points—a reminder that rejecting a null and measuring it precisely are different things.

A reader who cares about one specific outcome can combine these results with her own prior through the exchangeable-Bernoulli construction of Section 3.3. For the Peabody score—the borderline case, with unadjusted $p = 0.031$ but $\tilde{p}_{\mathrm{RW}} = 0.206$ in a family of $m = 7$—a development economist with prior $\pi = 0.3$ on a true effect updates to a posterior near 0.7. Within-family adjustment places Peabody outside the confirmatory rejection set, but its unadjusted evidence remains informative for a reader who brings a strong prior on the cognitive effects of better housing.

The exercise runs the framework end to end. The decision structure fixed the criterion (FWER, because the program decision is binary); the substantive logic of the program fixed the families and their causal ordering (HH_floor $\rightarrow$ CH_health $\rightarrow$ HH_satis); and the dependence within each family fixed the adjustment (the cluster bootstrap captures the cross-outcome correlation that analytical bounds miss), and the same resamples delivered the simultaneous intervals that report the magnitudes under that correction. The hierarchy holds here because the mechanism is correctly oriented—an analyst who ordered the families differently would have had to defend that ordering before seeing the data. Without a framework, any rejection set can be rationalized ex post by choosing the favorable method; with it, the method follows from the decision and the family structure, and the conclusion follows from the method.

7 Conclusion

The choice of error criterion in multiple testing follows from the structure of the decision the tests inform. When several tests feed into a binary action, even one false positive can lead to a wrong decision, and the familywise error rate is the natural object to control.

When the tests screen candidates for further investigation, the relevant quantity is the proportion of false leads, and the analysis should target the false discovery rate. The error criterion is not a methodological preference but a feature of the loss the researcher is trying to minimize, and the conventional significance level is itself the answer to a particular implicit cost-benefit calculation that a decision-maker facing a different trade-off may reasonably reject.

With the criterion fixed, the choice of procedure follows from the dependence structure of the data. Section 4 develops the case for Romano–Wolf as the default for confirmatory inference, Benjamini–Hochberg for exploratory work, hierarchical Romano–Wolf when the hypotheses come with a substantive ordering, and simultaneous confidence intervals as the magnitude analogue of adjusted tests. In each case the resampling-based or hierarchical procedure recovers power that the analytical bounds leave on the table.

None of these tools deliver their guarantees without pre-specification. The error rates that adjustment procedures control are statements about how often a fixed family of tests will produce a false rejection, and they do not survive selective reporting, omission of unfavorable hypotheses, or ex-post reordering of the testing sequence, whether or not the researcher intends those distortions. Pre-specifying the family, the error criterion, the procedure, and any hierarchical ordering before the data are examined converts the adjustment tools from ex-post rationalizations into valid inferential statements.

Two questions the framework raises remain open. The magnitude side of the problem has drawn far less attention than testing, even though in most applications the size of an effect matters as much as its significance: simultaneous confidence intervals (Section 4.5) are the natural tool, but interval estimation under multiplicity, and after data-driven model selection, is comparatively undeveloped. And how a reader who cares about one specific hypothesis should weigh it against the others tested remains, in the small families typical of economics, largely governed by her own prior (Section 3.3). Both are natural directions for further work.

Multiple testing adjustment should become as routine in applied economics as clustering standard errors — and the appropriate severity, like the choice to cluster, depends on the structure of the research environment rather than on statistical convention. The logic is the same in both cases. Ignoring a feature of the data-generating process leads to inference that misstates what the evidence supports, and the tools for doing better are well established and freely available. Pre-specifying the family, reporting adjusted p-values alongside unadjusted ones, and including simultaneous confidence intervals cost a researcher an afternoon. Not doing so produces conclusions that survive only until they fail to replicate.

REFERENCES

Anderson, Michael L., "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103* (484), 1481–1495.

_ **and Jeremy Magruder**, "Highly Powered Analysis Plans," Working Paper 29843, National Bureau of Economic Research 2022.

Andrews, Isaiah and Maximilian Kasy, "Identification of and Correction for Publication Bias," *American Economic Review*, 2019, *109* (8), 2766–2794.

Benjamini, Yoav and Daniel Yekutieli, "The Control of the False Discovery Rate in Multiple Testing under Dependency," *Annals of Statistics*, 2001, *29* (4), 1165–1188.

_ **and Yosef Hochberg**, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B*, 1995, *57* (1), 289–300.

Bretz, Frank, Willi Maurer, Werner Brannath, and Martin Posch, "A Graphical Approach to Sequentially Rejective Multiple Test Procedures," *Statistics in Medicine*, 2009, *28* (4), 586–604.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes, "Methods Matter: *p*-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 2020, *110* (11), 3634–3660.

Calonico, Sebastian, Sebastian Galiani, and Raul A. Sosa, "`mht`: Multiple-Hypothesis Testing in Stata," 2026. Working paper, companion to the present article.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Cattaneo, Matias D., Sebastian Galiani, Paul J. Gertler, Sebastian Martinez, and Rocio Titiunik, "Housing, Health, and Happiness," *American Economic Journal: Economic Policy*, 2009, *1* (1), 75–105.

Christensen, Garret and Edward Miguel, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 2018, *56* (3), 920–980.

Clarke, Damian, Joseph P. Romano, and Michael Wolf, "The Romano–Wolf Multiple Hypothesis Correction," *Stata Journal*, 2020, *20* (4), 812–843.

Dmitrienko, Alex, Ajit C. Tamhane, and Frank Bretz, eds, *Multiple Testing Problems in Pharmaceutical Statistics*, Boca Raton, FL: Chapman & Hall/CRC Press, 2009.

Efron, Bradley, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge: Cambridge University Press, 2010.

Gelman, Andrew and Eric Loken, "The Garden of Forking Paths," 2013. Unpublished manuscript, Department of Statistics, Columbia University.

Goldsmith-Pinkham, Paul, "Economics Literature Search," Online tool for searching empirical economics articles 2024.

Harvey, Campbell R., Alessio Sancetta, and Yuqian Zhao, "What Threshold Should Be Applied to Tests of Factor Models?," Working Paper 34898, National Bureau of Economic Research 2026.

Hochberg, Yosef, "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 1988, *75* (4), 800–802.

Holm, Sture, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 1979, *6* (2), 65–70.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

Lee, Jason D., Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor, "Exact Post-Selection Inference, with Application to the Lasso," *Annals of Statistics*, 2016, *44* (3), 907–927.

Lehmann, E. L. and Joseph P. Romano, "Generalizations of the Familywise Error Rate," *Annals of Statistics*, 2005, *33* (3), 1138–1154.

List, John A., Azeem M. Shaikh, and Yang Xu, "Multiple Hypothesis Testing in Experimental Economics," *Experimental Economics*, 2019, *22* (4), 773–793.

Olken, Benjamin A., "Promises and Perils of Pre-Analysis Plans," *Journal of Economic Perspectives*, 2015, *29* (3), 61–80.

Reif, Julian, "wyoung: Westfall–Young Multiple Testing Adjustments," Statistical Software Components S458854, Boston College Department of Economics 2024. Version 2.0, December 2024.

Romano, Joseph P. and Michael Wolf, "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 2005, *100* (469), 94–108.

_ and _, "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 2005, *73* (4), 1237–1282.

_ and _, "Balanced Control of Generalized Error Rates," *Annals of Statistics*, 2010, *38* (1), 598–633.

_ and _, "Efficient Computation of Adjusted *p*-Values for Resampling-Based Stepdown Multiple Testing," *Statistics & Probability Letters*, 2016, *113*, 38–40.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons, "*p*-Curve: A Key to the File-Drawer," *Journal of Experimental Psychology: General*, 2014, *143* (2), 534–547.

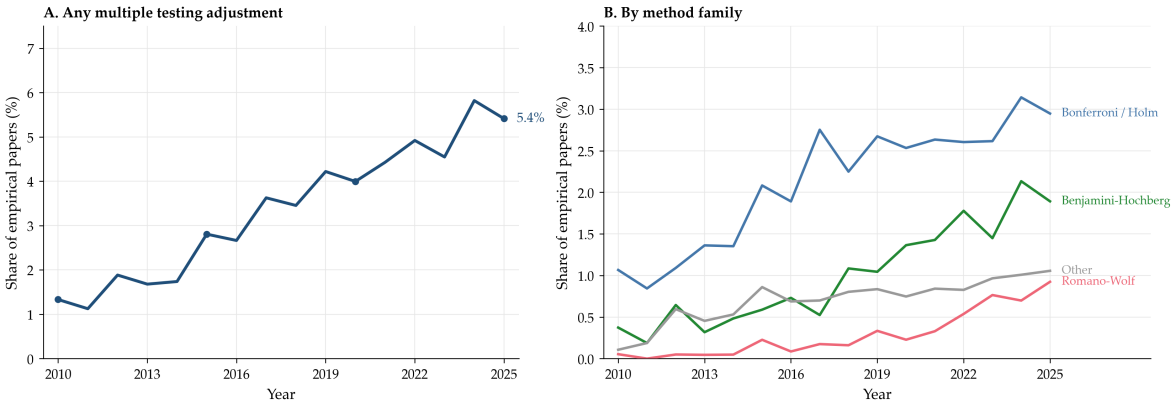
Storey, John D., "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society, Series B*, 2002, *64* (3), 479–498.

Viviano, Davide, Kaspar Wüthrich, and Paul Niehaus, "A Model of Multiple Hypothesis Testing," *Review of Economic Studies*, 2026. Forthcoming.

Westfall, Peter H. and S. Stanley Young, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, New York: Wiley, 1993.

Wiens, Brian L., "A Fixed Sequence Bonferroni Procedure for Testing Multiple Endpoints," *Pharmaceutical Statistics*, 2003, *2* (3), 211–215.

FIGURE 1. Adoption and Diversity of Multiple Testing Corrections in Economics, 2010–2025



Notes. Panel A reports the share of papers mentioning at least one formal multiple-testing term (Bonferroni, Šidák, Holm, false discovery rate, Benjamini–Hochberg, q-value, Romano–Wolf, Westfall–Young, familywise error, hierarchical/gatekeeping/fixed-sequence procedure); Panel B decomposes it into four method families, which are not mutually exclusive. Detection is by keyword search over the Goldsmith–Pinkham economics literature pipeline (Goldsmith-Pinkham, 2024) (full text for $\approx 87\%$ of papers). The corpus is 48,340 papers across fourteen outlets—AER, QJE, JPE, Econometrica, ReStud, the four AEJs, JF, JFE, RFS, JoE, and NBER Working Papers—indexed from 2010 onward. A keyword match captures *mention*, not necessarily *use*, which biases levels upward but not the trend.

TABLE 1. Software and Methods Reference

Panel A. Software implementations.

Language	Package	Methods	Notes
Stata	`mht` (Calonico et al., 2026)	Bonferroni, Šidák, Holm, Hochberg, BH, BY, RW (cluster-pairs or wild cluster bootstrap), hierarchical	Full §4 menu in one command; `mht hier` subcommand for hierarchical gatekeeping; cross-platform C plugin for the wild cluster path
Stata	`rwolf` (Clarke et al., 2020)	RW	OLS, IV, probit, clustered designs
Stata	`rwolf2`	RW	Successor to `rwolf`; supports multiple models, specifications, and weights
Stata	`wyoung` (Reif, 2024)	WY	Step-down resampling (Westfall and Young, 1993); supports strata and clustering
Stata	`mhtexp` / `mhtexp2`	FWER (experimental designs)	List et al. (2019) procedure; `mhtexp2` adds covariate adjustment
R	`p.adjust()` (base)	Bonferroni, Holm, Hochberg, Hommel, BH, BY	Operates on a p-value vector
R	`StepwiseTest`	RW; Hsu–Kuan–Yen FDP	Stepdown bootstrap implementation
R	`multtest` (Bioc.)	WY and related	Bootstrap and permutation MTPs; originally genomics-oriented
Python	`statsmodels.stats.multitest`	Bonferroni, Šidák, Holm, Hochberg, Hommel, BH, BY, two-stage variants	`multipletests()` on p-value array

Panel B. Summary of methods.

Method	Controls	Dependence	Computation	Recommended for	Avoid when
Bonferroni	FWER (strong)	Any	Trivial	Simple threshold	Holm is feasible
Šidák	FWER (strong)	Independence	Trivial	Rarely	Dependent tests
Holm	FWER (strong)	Any	Trivial	FWER fallback	Correlated outcomes (use RW instead)
Hochberg	FWER (strong)	PRDS	Trivial	Independent or PRDS	Negative dependence
BH	FDR	PRDS	Trivial	Default for FDR	Confirmatory settings
BY	FDR	Any	Trivial	Arbitrary depend. FDR control	Few tests (too conservative)
WY	FWER (strong)	Via permutation	Resampling	Experimental designs	Non-pivotal statistics
RW	FWER (strong)	Via bootstrap	Resampling	Default for FWER	Very small n

Notes. Panel A. Package capabilities evolve; consult documentation for current options. *Panel B.* PRDS = positive regression dependence on a subset. BH = Benjamini–Hochberg, BY = Benjamini–Yekutieli, WY = Westfall–Young, RW = Romano–Wolf.

TABLE 2. Simulation Results: Flat Dependence Designs

	$\rho = 0$			$\rho = 0.3$			$\rho = 0.7$		
Method	FWER	Power	FDR	FWER	Power	FDR	FWER	Power	FDR
Panel A: AR(1) dependence, sparse signal (20% non-null, $\delta = 0.25$)									
Unadjusted	0.528	0.709	0.177	0.518	0.722	0.174	0.396	0.710	0.159
Bonferroni	0.033	0.301	0.020	0.035	0.300	0.022	0.033	0.307	0.019
Šidák	0.033	0.303	0.020	0.035	0.304	0.022	0.035	0.310	0.020
Holm	0.037	0.306	0.021	0.035	0.308	0.022	0.038	0.314	0.020
Hochberg	0.037	0.306	0.021	0.035	0.308	0.022	0.038	0.314	0.020
RW	0.034	0.306	0.020	0.034	0.310	0.021	0.041	0.336	0.022
WY	0.034	0.306	0.020	0.034	0.310	0.021	0.041	0.336	0.022
BH	0.091	0.396	0.035	0.098	0.399	0.039	0.079	0.410	0.033
BY	0.019	0.221	0.012	0.019	0.231	0.012	0.019	0.244	0.009
Panel B: AR(1) dependence, dense signal (50% non-null, $\delta = 0.25$)									
Unadjusted	0.359	0.707	0.054	0.358	0.708	0.055	0.305	0.710	0.057
Bonferroni	0.018	0.298	0.005	0.024	0.293	0.008	0.025	0.290	0.010
Šidák	0.018	0.301	0.005	0.024	0.296	0.008	0.025	0.293	0.010
Holm	0.021	0.316	0.005	0.028	0.313	0.008	0.030	0.311	0.010
Hochberg	0.021	0.316	0.005	0.028	0.314	0.008	0.030	0.312	0.010
RW	0.021	0.315	0.005	0.029	0.313	0.009	0.034	0.335	0.011
WY	0.021	0.315	0.005	0.029	0.313	0.009	0.034	0.335	0.011
BH	0.133	0.520	0.022	0.134	0.515	0.022	0.111	0.515	0.023
BY	0.021	0.297	0.004	0.028	0.299	0.007	0.033	0.297	0.009
Panel C: Equicorrelated dependence, sparse signal (20% non-null, $\delta = 0.25$)									
Unadjusted	0.528	0.709	0.177	0.442	0.705	0.175	0.241	0.694	0.144
Bonferroni	0.033	0.301	0.020	0.022	0.296	0.014	0.012	0.288	0.007
Šidák	0.033	0.303	0.020	0.022	0.298	0.014	0.012	0.290	0.007
Holm	0.037	0.306	0.021	0.025	0.302	0.014	0.015	0.293	0.007
Hochberg	0.037	0.306	0.021	0.025	0.302	0.014	0.015	0.293	0.007
RW	0.034	0.306	0.020	0.026	0.316	0.014	0.030	0.393	0.020
WY	0.034	0.306	0.020	0.026	0.316	0.014	0.030	0.393	0.020
BH	0.091	0.396	0.035	0.063	0.396	0.025	0.032	0.390	0.015
BY	0.019	0.221	0.012	0.014	0.235	0.006	0.011	0.236	0.004
Panel D: Equicorrelated dependence, dense signal (50% non-null, $\delta = 0.25$)									
Unadjusted	0.359	0.707	0.054	0.333	0.704	0.064	0.189	0.694	0.088
Bonferroni	0.018	0.298	0.005	0.017	0.291	0.009	0.011	0.286	0.006
Šidák	0.018	0.301	0.005	0.017	0.293	0.008	0.011	0.289	0.006
Holm	0.021	0.316	0.005	0.020	0.311	0.008	0.015	0.311	0.007
Hochberg	0.021	0.316	0.005	0.020	0.311	0.008	0.015	0.312	0.007
RW	0.021	0.315	0.005	0.023	0.326	0.010	0.027	0.404	0.015
WY	0.021	0.315	0.005	0.023	0.326	0.010	0.027	0.404	0.015
BH	0.133	0.520	0.022	0.106	0.506	0.021	0.052	0.492	0.014
BY	0.021	0.297	0.004	0.016	0.303	0.004	0.016	0.308	0.004

Notes. Each entry is computed from 1,000 Monte Carlo replications with $n = 400$ observations and $m = 20$ hypotheses; treatment is $T_i \sim \text{Bernoulli}(0.5)$ and all procedures use $\alpha = 0.05$. Panels A–B use AR(1) covariance $\Sigma_{jk} = \rho^{|j-k|}$; Panels C–D use equicorrelated (compound-symmetric) covariance $\Sigma_{jk} = \rho$ for $j \neq k$, a single-factor model with uniform loadings. The sparse configuration has 20% of coefficients non-null at $\delta = 0.25$ and the dense configuration 50%; non-null positions are drawn uniformly from $\{1, \ldots, m\}$ and held fixed across correlation levels and structures via a shared seed. FWER is the fraction of replications with at least one false rejection; Power is the average fraction of non-null hypotheses correctly rejected; FDR is the average false discovery proportion $V / \max(R, 1)$. Global-null results are summarized in the text. The Romano–Wolf bootstrap uses $B = 1{,}000$ replications. BH = Benjamini–Hochberg, BY = Benjamini–Yekutieli, RW = Romano–Wolf, WY = Westfall–Young.

TABLE 3. Hierarchical Testing Results ($\rho = 0.5$, Three-Level Design)

Method	FWER	L1 power	L2 power	L3 power	Overall power	FDR
Unadjusted	0.408	0.976	0.700	0.326	0.637	0.081
Bonferroni	0.027	0.823	0.287	0.067	0.333	0.009
Šidák	0.028	0.827	0.291	0.067	0.335	0.009
Holm	0.032	0.835	0.302	0.073	0.344	0.010
Hochberg	0.032	0.835	0.303	0.073	0.344	0.010
Romano–Wolf	0.036	0.841	0.313	0.077	0.352	0.011
Westfall–Young	0.036	0.841	0.313	0.077	0.352	0.011
Benjamini–Hochberg	0.136	0.915	0.486	0.171	0.476	0.027
Benjamini–Yekutieli	0.035	0.797	0.294	0.070	0.331	0.009
Hierarchical Holm	0.056	0.974	0.409	0.080	0.425	0.012
Hierarchical RW	0.057	0.974	0.418	0.084	0.430	0.012

Notes. Three-level hierarchical design with $n = 400$ observations and $m = 20$ hypotheses. Level 1 (primary, two outcomes): strong effects $\delta_1 = 0.40$. Level 2 (mechanism, eight outcomes): four with moderate effects $\delta_2 = 0.25$, four null. Level 3 (exploratory, ten outcomes): three with weak effects $\delta_3 = 0.15$, seven null. Errors multivariate normal with AR(1) covariance at $\rho = 0.5$. Each entry is based on 1,000 Monte Carlo replications. Hierarchical Holm and Hierarchical RW use parent-specific propagation: a child is tested only if its parent rejects. Parent map: L2 outcomes 3–6 have L1 parent 1, 7–10 have parent 2; L3 outcomes 11–13 have L2 parents 3, 3, 4; L3 outcomes 14–20 have L2 parents 4 through 10.

TABLE 4. Multiple-Testing Adjustments Applied to the Piso Firme Evaluation

Outcome	$\hat{\beta}$ (SE)	Unadj. p	Within-family adjusted p			maxT 95% CI
			Holm	BH	RW	
Family 1: HH_floor (cement-floor mechanism, $m = 5$)						
Share of rooms with cement floor	0.210 (0.019)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{***}	[0.152, 0.269]
Cement floor in kitchen	0.265 (0.023)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{***}	[0.195, 0.335]
Cement floor in dining room	0.221 (0.025)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{***}	[0.145, 0.297]
Cement floor in bathroom	0.117 (0.018)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{***}	[0.062, 0.171]
Cement floor in bedroom	0.245 (0.020)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{***}	[0.184, 0.306]
Family 2: CH_health (child-health target, $m = 7$)						
Parasitic infestation count	−0.064 (0.032)	0.046^{*}	0.186	0.076	0.236	[−0.158, 0.031]
Diarrhea (last 30 days)	−0.018 (0.009)	0.054	0.186	0.076	0.236	[−0.046, 0.010]
Anemia	−0.083 (0.027)	0.002^{**}	0.015^{*}	0.009^{**}	0.037^{*}	[−0.162, −0.003]
MacArthur communicative development	5.557 (1.641)	0.001^{***}	0.007^{**}	0.007^{**}	0.020^{*}	[0.671, 10.442]
Peabody vocabulary score	3.083 (1.410)	0.031^{*}	0.153	0.071	0.206	[−1.113, 7.279]
Height-for-age (z-score)	0.002 (0.039)	0.960	1.000	0.960	0.951	[−0.114, 0.117]
Weight-for-height (z-score)	−0.011 (0.037)	0.766	1.000	0.894	0.944	[−0.120, 0.098]
Family 3: HH_satis (downstream wellbeing, $m = 5$)						
Satisfaction with floor quality	0.222 (0.026)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{***}	[0.152, 0.292]
Satisfaction with house quality	0.084 (0.022)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.002^{**}	[0.024, 0.145]
Satisfaction with quality of life	0.112 (0.022)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{***}	[0.052, 0.172]
CES-D depression scale	−2.372 (0.562)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{***}	[−3.897, −0.847]
Perceived stress scale	−1.742 (0.396)	$<0.001^{***}$	$<0.001^{***}$	$<0.001^{***}$	0.001^{***}	[−2.816, −0.669]
Rejections at $\alpha = 0.05$	—	14	12	12	12	12 excl. 0

Notes. OLS regressions of each Piso Firme outcome on the treatment indicator with the Model 3 controls of Cattaneo et al. (2009); HH_floor and HH_satis outcomes are at the household level ($n \approx 2{,}755$), CH_health outcomes at the child level (n varies 593–4,051). Cluster-robust standard errors at the geographic-cluster level (113–136 clusters). Each pre-specified family is adjusted on its own: Holm, Benjamini–Hochberg, and Romano–Wolf with a cluster bootstrap of $B = 1{,}000$ draws, via the Stata command `mht` (Calonico et al., 2026). Stars denote rejection at $\alpha = 0.05$ (*), 0.01 (**), and 0.001 (***); the resampling granularity floor is $1/(B+1) \approx 0.001$. The final column is the maxT 95% simultaneous confidence interval $[\hat{\beta}_j \pm c\,\mathrm{SE}_j]$ from the same bootstrap, with family-specific critical values $c \approx 3.04$ (HH_floor), 2.98 (CH_health), and 2.71 (HH_satis); by the duality of Section 4.5 a band excludes zero exactly when the Romano–Wolf test rejects, so twelve of the seventeen bands exclude zero. The hierarchical reading (Section 4.4) orders the families HH_floor → CH_health → HH_satis and tests each level only if the prior family rejects at least one outcome; here all three gates open, so the hierarchical verdict coincides with the per-family Romano–Wolf column.