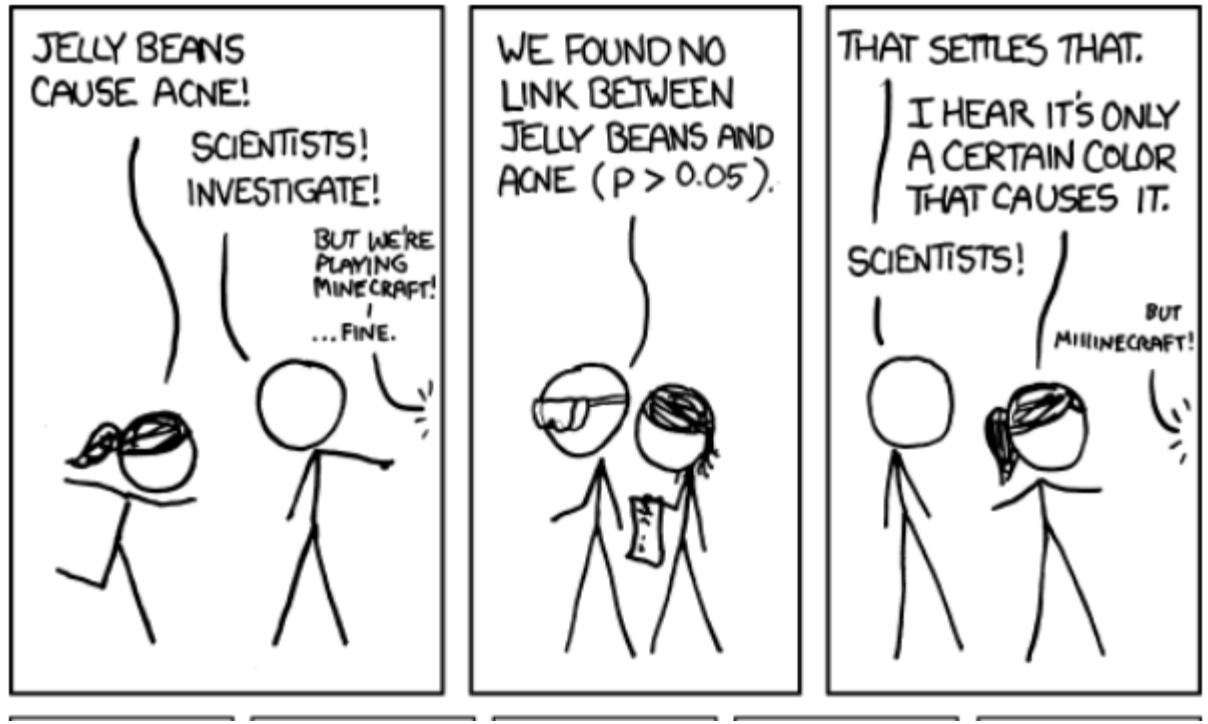
Multiple testing, pre-analysis plans, and registered reports

David McKenzie, *World Bank*

Significant

|< < Prev Random Next > >|




WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Common to see tables with lots of tests: e.g. multiple treatments (4) * multiple outcomes (10) = 40 tests here.

Appendix Table 5.2: Practice by Practice Impacts on Finance & Accounting Practices in Round 2

	Records money in	Records money out	Income Statement	Balance Sheet	Cash Flow Statement	Knows most profitable	Uses cost-control	Prepared a Budget	Sets Financial Goals	Conducts Feasibility Studies
Assigned to Insourcing	0.105**	0.077	0.098	0.008	0.024	0.052	0.047	0.019	-0.035	0.037
	(0.053)	(0.055)	(0.060)	(0.060)	(0.056)	(0.048)	(0.051)	(0.060)	(0.054)	(0.055)
Assigned to Outsourcing	0.071	0.081	0.047	0.084	0.046	0.040	0.079	-0.010	0.076	0.084
	(0.054)	(0.055)	(0.060)	(0.060)	(0.056)	(0.048)	(0.049)	(0.059)	(0.050)	(0.054)
Assigned to Training	0.086	0.073	0.018	0.025	-0.042	-0.026	0.011	-0.047	-0.041	-0.117**
	(0.054)	(0.056)	(0.061)	(0.060)	(0.058)	(0.051)	(0.053)	(0.061)	(0.054)	(0.058)
Assigned to Consulting	0.150***	0.151***	0.114*	0.110*	0.059	0.060	0.067	0.058	0.059	0.101*
	(0.051)	(0.053)	(0.060)	(0.060)	(0.055)	(0.046)	(0.049)	(0.059)	(0.050)	(0.053)
Mean of Control Group	0.682	0.659	0.447	0.424	0.674	0.788	0.758	0.492	0.750	0.674
Sample Size	678	678	678	678	678	678	678	678	678	678
P-value: all treatments zero	0.059	0.081	0.243	0.259	0.391	0.334	0.419	0.502	0.059	0.001
P-value: all treatments equal	0.367	0.322	0.348	0.272	0.283	0.296	0.532	0.345	0.033	0.000
P-value: In=Out=Consult	0.253	0.237	0.508	0.205	0.810	0.907	0.792	0.500	0.070	0.463
P-value: Insource = Outsource	0.504	0.944	0.395	0.202	0.691	0.792	0.496	0.616	0.028	0.373

Also a very common issue with heterogeneity analysis

Table 7
Heterogeneous impact on formalization by baseline characteristics.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variables:	Formalized: GUFE data						
	Female owner	Operates in Dantokpa market	Trader	Doesn't look like formal species	Index of business size below median	Does not have secondary education	One visit or fewer from tax inspectors
Variable for heterogeneous analysis:							
Impact in group [...] for heterogeneous variable = 0							
Group1	0.134***	0.105***	0.144***	0.125**	0.085***	0.140***	0.124**
	(0.035)	(0.026)	(0.032)	(0.055)	(0.032)	(0.036)	(0.054)
Group2	0.192***	0.151***	0.178***	0.224***	0.139***	0.175***	0.176***
	(0.024)	(0.016)	(0.021)	(0.035)	(0.020)	(0.024)	(0.036)
Group3	0.206***	0.179***	0.195***	0.231***	0.151***	0.218***	0.214***
	(0.021)	(0.014)	(0.019)	(0.032)	(0.018)	(0.022)	(0.033)
Additional impact in group [...] for heterogeneous variable = 1							
Group1 × Heterogenous variable (int1)	− 0.063	− 0.048	− 0.089**	− 0.036	0.022	− 0.074	− 0.035
	(0.046)	(0.054)	(0.045)	(0.061)	(0.046)	(0.049)	(0.061)
Group2 × Heterogenous variable (int2)	− 0.096***	− 0.100***	− 0.086***	− 0.115***	− 0.017	− 0.073**	− 0.056
	(0.029)	(0.034)	(0.028)	(0.039)	(0.029)	(0.033)	(0.041)
Group3 × Heterogenous variable (int3)	− 0.070***	− 0.080***	− 0.058**	− 0.083**	0.022	− 0.096***	− 0.064*
	(0.026)	(0.031)	(0.025)	(0.036)	(0.026)	(0.031)	(0.038)
Observations	3596	3596	3596	3596	3596	3596	3596
R-squared	0.395	0.395	0.396	0.394	0.392	0.400	0.393
Adjusted R-squared	0.090	0.089	0.090	0.088	0.085	0.097	0.086
Mean heterogenous variable	0.629	0.217	0.550	0.818	0.500	0.591	0.804

Young (2019)

- I examine 53 papers, 14 of which are laboratory experiments and 39 of which are field experiments. 27 of the papers appeared in the AER, 21 in the AEJ: Applied 9 Economics, and 5 in the AEJ: Microeconomics
- The number of tables reporting estimates and standard errors for treatment effects …varies substantially across papers, with 17 papers having only 1 or 2 such tables and 19 presenting 5 to 8.
- The number of treatment effects reported in tables ranges from 2 to 96, with the average table in a paper having a mean of 19 and median of 17
- when a table reports a .01 significant result, on average there are 21.2 reported treatment effects and only 5.0 of these are significant

Outline for today

- **Different approaches for handling multiple hypothesis testing and when to use each?**
  - Aggregate index approach
  - Controlling the FWER
  - Controlling the FDR
  - Omnibus testing/Joint significance approach
- **Enhancing credibility and tying one's hands through pre-committing to which tests you will run**
  - Pre-registration of studies: AEA RCT registry and options for non-RCTs
  - Pre-analysis plans
  - Registered reports

The Aggregate Index Approach

- *Problem:* too many hypotheses being tested due to lots of outcomes
- *Solution:* collapse the number of outcomes into a single summary measure
- *Example:* I'm interested in seeing whether business training improves accounting practices in firms, and measure 10 different practices.
  - Rather than focusing on impacts one by one on each of these 10 practices, form an overall measure of accounting practices and look at impact on this.

  - *Note second advantage of these measures – can boost statistical power – e.g. get marginally significant impacts on range of proxies, when aggregate together get rid of noise and more significant.*

How should we aggregate outcomes into an index?

- Approach one: simple average of binary variables
  - E.g. set of measures of whether firm is doing business practice X; or set of health behaviors during Covid, etc.
  - Then just take the mean of all these binary variables.
  - Outcome is then interpretable as the proportion of these practices the firm is doing.
  - Potential downside: puts equal weight on every outcome.
- Approach two: first principal component
  - Common technique for dimensionality reduction
  - Famous example: Filmer and Pritchett with asset indicators
  - Don’t need all the variables to be binary
  - Units harder to interpret

How should we aggregate outcomes into an index?

- Approach three: average of normalized z-scores (Kling et al, 2007)
  - Organize variables so all go in same direction. E.g. suppose want to see if cash transfers improve child health, and have outcomes
    - BMI
    - Number of episodes of illness in last 2 weeks (need to reverse code e.g. number of days without illness)
    - Child height
    - Anemia (need to reverse code e.g. not anemic)
  - Normalize each outcome as a z-score by subtracting control mean and dividing by control s.d. e.g. $z_{BMI,i} = (BMI_i - \overline{BMI}_{control})/\text{sd}_{\text{control}}$
  - Then take simple average of these z-scores
  - Coefficient typically interpreted as standard deviation change in index (*but note two concerns with s.d. units*)

How should we aggregate outcomes into an index?

- Approach four: inverse-covariance weighted (Anderson, 2008; O'Brien, 1984)
  - Rather than taking simple average of the z-scores, use inverse-covariance weights
  - ensures that outcomes that are highly correlated with each other receive less weight, while outcomes that are uncorrelated and thus represent new information receive more weight
  - Samii (2016) gives example of college math grade, math GRE, verbal GRE – "The inverse covariance weighted average of these three variables would result in an index that gives about 25% weight to each math score and then 50% weight to the verbal score. It "rewards" the verbal score for providing new information that the math scores don't. The resulting index could be interpreted as a "general scholastic aptitude" index."

Aggregate indices in practice

- What variables can I aggregate together?
  - E.g. does it make sense to put employment, teen pregnancy, household consumption, agricultural productivity and tv watching together?
  - Usually want a *domain* of behavior/meaningful sense of an underlying concept
  - E.g.2: what about firm survival, profits, sales, and number of workers?
  - Do not want a bunch of junk variables in there
- Thinking carefully about using variables with no variation
  - E.g. business practices, if have practices everyone does or no one does
- Does it matter which aggregation method I use?
  - Sometimes, but in practice I typically find correlations are >0.99. But depends on how you form the index/types of variables you put in
- What should you do if missing values for some components
  - E.g. income missing for 10% of individuals, employment for 3%, occupation there for everyone.

Adjust p-values for multiple inference by controlling the FWER

- **Family-wise error rate (FWER)**
  - Suppose I test *M* hypotheses, H_1, H_2, …, H_M, of which J are true
  - The FWER is the probability that at least one of the J true hypotheses in the family is rejected
  - M=1, then FWER = type-I error rate (e.g. 5%) – incorrectly reject a true null (*get a false positive)*
  - As M grows, the chance of rejecting at least one hypothesis at a given significance level α increases, and FWER grows.
    - E.g. 2 independent tests $1-0.95^2 = 0.098$
      - 5 independent tests $1-0.95^5 = 0.226$
      - 10 independent tests $1-0.95^{10} = 0.401$
      - 20 independent tests $1-0.95^{20} = 0.641$

Controlling the FWER

- **Simplest approach**: Bonferroni adjustment
  - With M tests, reject the null only if $p \leq \frac{\alpha}{M}$
  - Equivalently, adjusted p-value is $p_{Bonferroni} = min[Mp, 1]$

- Simple – only need to know p-value and number of tests
- But usually far too conservative, increasing risk of type II errors (failing to reject null when null is false)
- Does not account for correlation among outcomes

Controlling the FWER

- In practice, better to use *Westfall-Young* or *Romano-Wolf*.
  - These are step-down resampling methods – resample data lots of time, preserving correlations among outcomes. This can greatly increase power if outcomes highly correlated.
  - When a hypothesis is rejected, step-down resampling method removes it from family being tested, increasing power for the remaining tests.
  - In Stata: *rwolf2* and *wyoung* commands

  https://blogs.worldbank.org/impactevaluations/updated-overview-multiple-hypothesis-testing-commands-stata

Example from Anderson (2008) with 9 outcomes (each a summary index)

Table 3. Summary index effects

		Female					Male				Gender
Project	Age	Effect	Naive p value	FWER p value	n	Bonferroni	Effect	Naive p value	FWER p value	n	difference t statistic
ABC	Preteen	.445	.026	.125	54	0.234	.417	.026	.184	51	.11
		(.194)					(.181)				
Perry	Preteen	.537	.004	.028	51	0.036	.150	.387	.943	72	1.53
		(.177)					(.172)				
ETP	Preteen	.362	.160	.349	30	1.00	.148	.552	.958	34	.61
		(.251)					(.245)				
ABC	Teen	.422	.042	.156	53	0.38	.162	.407	.943	51	.93
		(.202)					(.194)				
Perry	Teen	.613	0	.003	51		.035	.716	.977	72	3.32
		(.156)					(.096)				
ETP	Teen	.456	.138	.349	29		.123	.747	.977	32	.68
		(.299)					(.377)				
ABC	Adult	.452	.003	.024	53		.312	.066	.372	51	.64
		(.144)					(.166)				
Perry	Adult	.353	.022	.125	51		−.012	.927	.977	72	1.83
		(.150)					(.130)				
ETP	Adult	−.069	.714	.701	29		−.710	.011	.090	31	1.98
		(.186)					(.260)				

NOTE: Parentheses contain OLS standard errors. Naive p values are unadjusted p values based on the t distribution. FWER p values adjust for multiple testing at the summary index level and are computed as described in Section 3.2.2. The t statistics test the difference between female and male treatment effects. See Table 2 for the components of each summary index.

Adjust p-values for multiple inference by controlling the FDR

- **FWER** approach is appropriate when you want to guard against **any** false positives. This may be appropriate when cost of a false rejection is high (e.g. large scaling up of ineffective program). But the more tests you do, the more restrictive it becomes.
- Alternatively, we may be willing to tolerate some type I errors in exchange for more power, especially in exploratory analysis.
- The **False discovery rate** (FDR) is designed to control the proportion of false positives among the set of rejected hypotheses.
  - When some false hypotheses are correctly rejected, FDR is less than FWER, and need less stringent adjustment to p-values.
  - Can even get cases where, by rejecting a lot of hypotheses, adjusted p-values are less than unadjusted.

FDR adjustment in practice

- Benjamini and Hochberg (1995), as implemented by Anderson (2008), gives *sharpened q-values*
- Doesn't account for correlation, fine for independent or positive, but may need to adjust if negative correlations

Example of Five Outcomes and Four Treatments, with sharpened q-values

	Y1	Y2	Y3	Y4	Y5
Treat 1	0.022	0.043	0.083**	0.079***	0.032
p-value	(0.516)	(0.258)	(0.031)	(0.001)	(0.178)
sharpened q-value	[0.381]	[0.255]	[0.091]	[0.011]	[0.179]
Treat 2	0.043	0.060	0.099***	0.083***	0.046**
p-value	(0.168)	(0.109)	(0.006)	(0.001)	(0.048)
sharpened q-value	[0.179]	[0.151]	[0.038]	[0.011]	[0.109]
Treat 3	0.030	-0.006	-0.016	0.008	0.009
p-value	(0.356)	(0.877)	(0.665)	(0.726)	(0.691)
sharpened q-value	[0.312]	[0.443]	[0.381]	[0.381]	[0.381]
Treat 4	0.042	0.093**	0.070*	0.044*	0.052**
p-value	(0.179)	(0.014)	(0.052)	(0.064)	(0.024)
sharpened q-value	[0.179]	[0.064]	[0.109]	[0.116]	[0.084]
Sample Size	726	678	678	678	678

Omnibus testing/joint significance

- **Multiple treatments**: test the null hypothesis that all treatment effects jointly zero on an outcome (standard F-test)
- **Young test of complete irrelevance:** under the sharp null that no treatment has any impact for any unit on any outcome, can use permutation testing. *Randcmd* in Stata

Example of Five Outcomes and Four Treatments, with randcmd randomization-t p-values

	Y1	Y2	Y3	Y4	Y5	All equations
Treat 1	0.022	0.043	0.083**	0.079***	0.032	
p-value	(0.516)	(0.258)	(0.031)	(0.001)	(0.178)	
Treat 2	0.043	0.060	0.099***	0.083***	0.046**	
p-value	(0.168)	(0.109)	(0.006)	(0.001)	(0.048)	
Treat 3	0.030	-0.006	-0.016	0.008	0.009	
p-value	(0.356)	(0.877)	(0.665)	(0.726)	(0.691)	
Treat 4	0.042	0.093**	0.070*	0.044*	0.052**	
p-value	(0.179)	(0.014)	(0.052)	(0.064)	(0.024)	
Young Westfall-Young joint test	0.403	0.045	0.03	0.005	0.091	0.022
Sample Size	726	678	678	678	678	

Considerations for using in practice

- Separate outcomes into different *families* or *domains*. E.g. in moving-to-opportunity, might want to look at effect of better neighborhood in outcomes related to:
  - Economic self-sufficiency (work)
  - Mental health
  - Physical health
  - Risky behavior
  - Education
- May then want to form a summary index for each domain, and then do a FWER adjustment for looking at overall effect on these 5 indices.
- Then might want to separately look at impacts on individual outcomes within domain (maybe in appendix), and do FWER or FDR adjustment.
- With treatment heterogeneity, perhaps more exploratory, and maybe FDR adjustment.

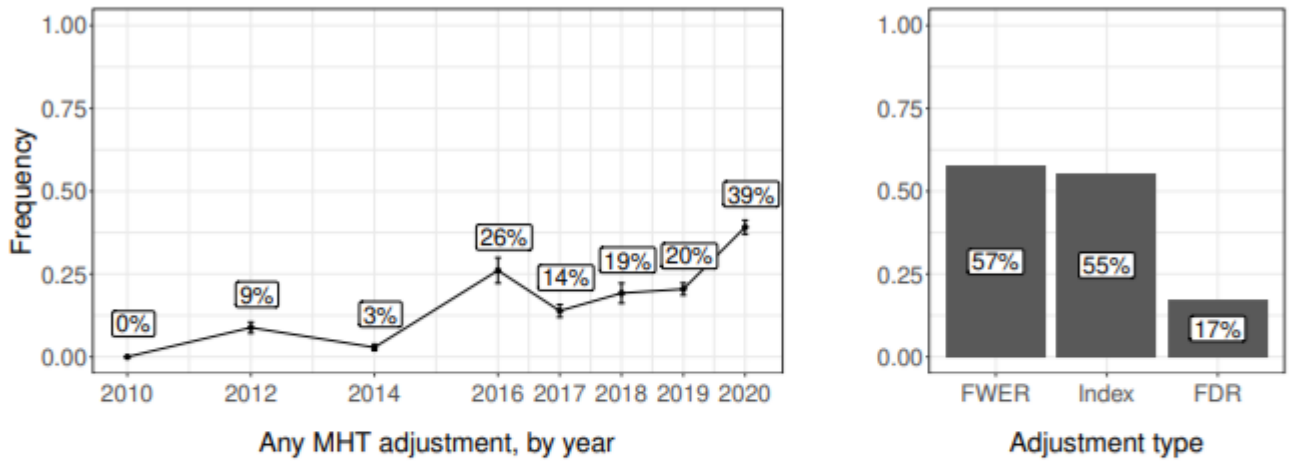
Do we even need to adjust at all?

- Depends on how we want to use results.
  - E.g. 1 Gibson et al. (2011) – Omnibus testing of migration's effects on remaining household members – 62 different outcomes.
  - If we want to compare impact of migration on business ownership from this study to those in other studies in the literature, don't need to adjust (since we are ignoring the other results)
  - Index measures may not be appropriate – e.g. looking at household income, we may be more interested in how composition changed (more remittances, less ag income, what happened to business income) then overall impact on income -> then FWER type adjustment.

Do we even need to adjust at all?

- Viviano et al. (2021)

Figure 1: Multiple hypothesis testing adjustment in "top-5" experimental publications

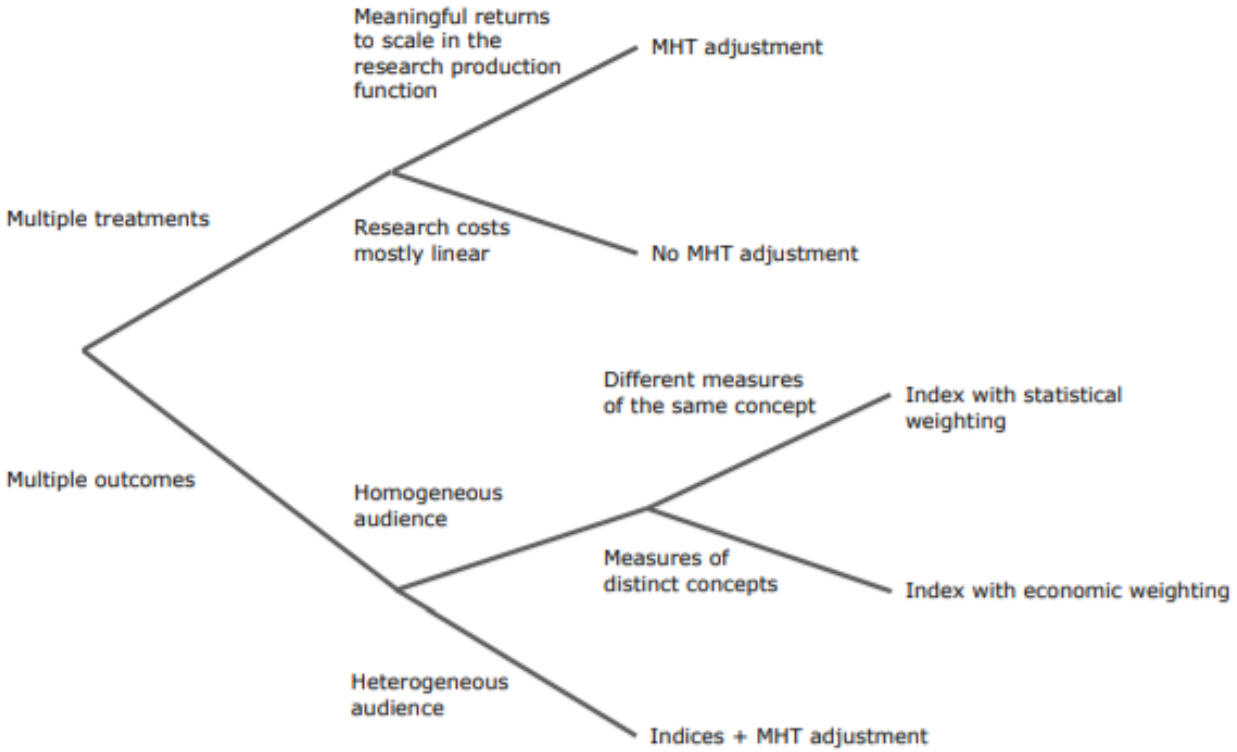


Notes: The left-hand panel reports the share of experimental papers that report at least one multiple hypothesis testing adjustment, including both indexing and control of compound error rates, by year of publication. The right-hand panel reports the frequency of each adjustment type, pooling across years. Adjustment types are not mutually exclusive. Authors' calculations based on a review of papers published in the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*.

Do we even need to adjust at all?

- Viviano et al. (2021) – depends on how costly it is to do extra tests (e.g. adding more interventions costly, more subgroup testing not); and on policy decision
- E.g. if policymaker cares about education, work, & health outcomes of program, get weights they assign to each to form overall index

Figure 3: Stylized summary of implications for practice



Enhancing credibility and tying one's hands through pre-committing to which tests you will run

- A more nefarious way to reduce the number of hypotheses that are being corrected is for researchers to not report every test they run:
  - Do a lot of "exploratory analysis", and then report the interesting things you find
  - P-hacking: explicitly test different ways of coding variable, different hypotheses, and only report those that show up significant

  $\Rightarrow$Get that Green jelly beans lead to acne headline
- Even with methods we've just discussed, lots of researcher decisions along the way:
  - Which outcomes get made into an index?
  - How will the index be formed?
  - How many families/domains?
  - Which method of adjustment for multiple testing?
  - Etc.

Tools for pre-commitment/pre-planning

- **Pre-registration of a study** – say what you are going to do before you see the data
- **Pre-analysis plan** – specify very precisely how you are going to the analysis
- **Registered report** – also get it peer-reviewed, write it up like a paper

Pre-registration

- For randomized experiments, main registry is the AEA RCT registry.
  - Mandatory now for submission to some journals (e.g. AEA journals:

> The American Economic Association operates a Registry for Randomized Controlled Trials (RCTs). As of January 2018, registration in the RCT registry is mandatory for all applicable submissions. This applies to field experiments. Laboratory experiments do not need to be registered at this time. You will be asked to provide your AEARCT identification number in the online submission form. Please include your number in the acknowledgement footnote in your paper, as well.

  - Just requires basic information in mandatory fields:
    - Dates
    - Intervention details
    - Main outcomes – can be specified with more or less precision, up to authors
  - Nearly 5,500 registrations with locations in 162 countries (Feb 2022)
  - Note: https://clinicaltrials.gov/ (required for medical journals)

INTERVENTIONS

Intervention(s)

Firms will be offered a mixture of information, training and assistance, and take-up incentives to encourage their take-up and usage of the new technology of digital marketing.

Experimental Design

Firms will be in one of four arms:

(1) Digital Marketing Training + Advertising Subsidy + Business Support
(2) Pay for Performance (Digital Marketing Presence and Advertising)
(3) Information Only
(4) Control

EXPERIMENT CHARACTERISTICS

Sample size: planned number of clusters

2550 firms

Sample size: planned number of observations

2550 small firms

Sample size (or number of clusters) by treatment arms

850 in training & subsidy, 850 in pay for performance, 300 in information, 550 in pure control

PRIMARY OUTCOMES

Primary Outcomes (end points)

We are interested in primary outcomes in two key domains:
Domain 1: Take-up and Usage of Digital Marketing Technology.
This will be measured by:
1. Percentage of firms with a Facebook page for the firm
2. Percentage of firms that have done paid advertising on Facebook in the last three months.

Domain 2: Firm performance
This will be measured by:
1. Sales made to new customers who have heard of firm through digital marketing efforts
2. Sales
3. Profits

Primary Outcomes (explanation)

These primary outcomes will be refined prior to collection of our first follow-up data, and will be explained in the pre-analysis plan

SECONDARY OUTCOMES

Secondary Outcomes (end points)

To be defined in pre-analysis plan

Specifying primary outcomes

- We often have lots of things we'd like to look at
  - Many different outcomes
  - Lots of different theories about heterogeneity
- I like to give myself and my co-authors what I call the Science/AER Insights/Policy brief test
  - Ask yourself – if at the end of this study, I was only allowed to show one short table or figure, what would be in it?
  - If you ask policymakers what outcomes matter most for them in making decisions based on this study, what would they say (and are you measuring these)?
- After all, you are not tying your hands very much if you say I intend to measure treatment effects on these 148 outcomes, and don't say whether some matter more than others.

Table 1. Impact of training programs on business survival, profitability, and sales. Data are from four rounds of surveys and show the average impacts over the 2.5 years after training. All regressions include randomization strata and survey wave dummies. Huber-White robust standard errors (in parentheses) are clustered at the firm level. Business survival is a binary indicator that takes the value 1 if the business survives. Sales are winsorized (capped) at the 99th percentile and profits at the 1st and 99th percentiles, reducing the influence of outliers. Sales and profits are expressed in terms of real CFA francs. The profits and sales index is the mean of the standardized *z*-scores of our various profits and sales measures. An *F* test was used to test equality of the impacts of the two training programs. *$P < 0.1$; **$P < 0.05$; ***$P < 0.01$.

	Business survival	Monthly sales	Monthly profits	Weekly profits	Profits and sales index
Traditional business training	−0.005	38,077	10,746	3086	0.029
	(0.008)	(57,812)	(6802)	(2050)	(0.030)
Personal initiative training	−0.003	114,733*	28,709***	6685***	0.100***
	(0.008)	(58,619)	(7110)	(1979)	(0.031)
Number of observations	5792	5642	5642	5633	5643
Number of firms	1499	1492	1492	1492	1492
P value from test of equality of treatments	0.813	0.171	0.014	0.091	0.025
Control group mean	0.960	680,807	96,089	30,417	0.000

Table 2. Mechanisms through which training operates. Huber-White robust standard errors (in parentheses) are clustered at the firm level. *$P < 0.1$; **$P < 0.05$; ***$P < 0.01$.

	Business practices	Personal initiative	Capital and labor inputs	Innovation index	Diversified product line	Access to finance index
Traditional business training	0.060*** (0.008)	0.065*** (0.015)	0.032* (0.020)	0.117*** (0.050)	0.044** (0.018)	0.070** (0.033)
Personal initiative training	0.054*** (0.007)	0.124*** (0.015)	0.078*** (0.020)	0.309*** (0.070)	0.092*** (0.018)	0.147*** (0.040)
Number of observations	5646	5538	5655	5639	5632	4207
Number of firms	1492	1484	1492	1492	1492	1473
P value from test of equality of treatments	0.458	0.000	0.024	0.011	0.010	0.043
Control group mean	0.618	4.32	0.000	0.000	0.335	0.000

Doesn’t mean you can’t look in quite a few domains, but need to be judicious – ultrapoor example

Table 3. Indexed family outcome variables and aggregates.

	Endline 1			Endline 2		
	(1)	(2)	(3)	(4)	(5)	(6)
Indexed outcomes	Standardized mean treatment effect	*q*-value for all 10 hypotheses	*F*-test of equality of coefficients across sites, with *q*-values	Standardized mean treatment effect	*q*-value for all 10 hypotheses	*F*-test of equality of coefficients across sites, with *q*-values
Total per capita consumption, standardized	0.122*** (0.023)	0.001	3.207 0.009	0.120*** (0.024)	0.001	5.307 0.001
Food security index (five components)	0.107*** (0.022)	0.001	1.670 0.139	0.113*** (0.022)	0.001	2.405 0.050
Asset index	0.258*** (0.023)	0.001	14.26 0.001	0.249*** (0.024)	0.001	23.90 0.001
Financial inclusion index (four components)	0.367*** (0.030)	0.001	55.33 0.001	0.212*** (0.031)	0.001	10.70 0.001
Total time spent working, standardized	0.090*** (0.018)	0.001	7.520 0.001	0.054*** (0.018)	0.004	2.644 0.038
Incomes and revenues index (five components)	0.383*** (0.036)	0.001	12.05 0.001	0.273*** (0.029)	0.001	5.82 0.001
Physical health index (three components)	0.034* (0.019)	0.078	3.825 0.003	0.029 (0.020)	0.159	0.776 0.630
Mental health index (three components)	0.099*** (0.022)	0.001	5.189 0.001	0.071*** (0.020)	0.001	1.781 0.142
Political Involvement index (four components)	0.064*** (0.018)	0.001	4.176 0.002	0.064*** (0.019)	0.002	2.624 0.038
Women's empowerment index (five components)	0.046** (0.023)	0.049	1.803 0.121	0.022 (0.025)	0.385	0.469 0.800

What about sub-groups/heterogeneity?

- Useful to pre-specify if one or two dimensions that theory/policy strongly focused on.
- But typically lots of potential dimensions of heterogeneity that might apply
  - Which are of most interest depends on whether main purpose is policy targeting or understanding economic mechanism
    - E.g. heterogeneity by geographic region, age group, gender vs
      heterogeneity by risk aversion, level of credit constraints, baseline level of capital
- Most experiments have low power for looking at heterogeneity
  - Consider as exploratory only
  - Use machine learning methods to look at heterogeneity over many variables, rather than testing lots of bivariate hypotheses.

What about non-experimental studies?

- Only makes sense if you can credibly commit to not having seen the data when you specify analysis.
- Examples:
  - Policy change (new minimum wage, new benefit) coming in, and specify how will use government survey data to be collected after the change.
  - Prospective impact evaluation:
    - E.g. regression-discontinuity design of some policy, and now planning follow-up survey to see what impacts are
    - E.g. prospective difference-in-differences design or synthetic control design – might pre-specify now how comparison groups be formed, and follow-up data not yet collected.
- Main place to register: Open science foundation (osf) (now includes EGAP)
- Alternative: Aspredicted.org

Pre-analysis plans

- A **pre-analysis plan** is a step-by-step plan setting out how a researcher will analyze data which is written in advance of them seeing this data (and ideally before collecting it in cases where the researcher is collecting the data).
- Typically more detailed than AEA registration:
  - Specify how data will be cleaned, exact regression specifications, exactly how variables will be constructed, etc.
- Some debate in the literature as to when/whether to do one and how extensive it should be:
  - Early pre-analysis plans often were 30+ pages, tried to pre-specify all eventualities
  - Olken (2005) suggests simpler plans may have most of advantages
  - Coffman and Niederle (2005) – may not be needed if easy to replicate – e.g. lab experiments.
  - Duflo et al. (2020) call for moderation

Pros of pre-analysis plans/when to use

- Most useful for field experiments that may be expensive and difficult to replicate
- As well as credibility, several other uses:
  - Helps focus policymakers on what key outcomes they most care about, and get agreement on this in advance rather than after seeing results.
  - Really helpful if you are designing questionnaires – helps make sure you are measuring everything you need to, and also on what can be cut from questionnaire if necessary (lots of “nice to ask” questions never get used).
  - Records upfront a lot of design and intervention details that may be harder to remember/reconstruct 2-3 years later.
  - Upfront investment in thinking through analysis can make it much faster/easier when data arrives to get headline results

What should go into a pre-analysis plan?

- See checklist: https://blogs.worldbank.org/impactevaluations/a-pre-analysis-plan-checklist

1) **Describe how sample is selected, expected sample size, how randomization is done**
2) **Key data sources and timing of data collection**
3) **Estimating equation:** e.g. what controls will be used in regression, how will standard errors be calculated, what adjustments will be used for multiple testing, etc.
4) **How will attrition be handled?**

What should go into a pre-analysis plan?

5) Hypotheses, families and outcomes:

- I see this as the most important.
- Separate outcomes into separate families/domains, and ideally also into primary and secondary outcomes, or main effects, mechanisms and heterogeneity.
- For outcomes, be really specific – so RA could create from these instructions:
  - E.g. don't just say "Wage earnings in last month", but make clear
    - Log, Levels, or some other transform? How will 0s be handled?
    - Any winsorizing to deal with outliers?
    - If you have multiple questions on this (e.g. earnings in main job, earnings in past week, etc) how will overall measure be constructed.
  - If forming summary index measures, define components of this and how will be constructed.

Example 1:

- YouWin business plan

FAMILY D: CHANGES IN BUSINESS SALES AND PROFITABILITY

HYPOTHESIS D1: Treatment leads to greater sales and profits in the medium term, but likely has no discernible impact in the first follow-up survey.

This will be measured as the following set of outcomes:

1. **Number of customers in a typical week** (B12). This will be top-coded at the 99th percentile of the overall distribution to account for outliers.
2. **Total sales in the last month with no truncation:** BF5. For businesses not answering the exact answer, but answer the range question, the midpoint of the range will be used. For firms in the top range, a value equal to the median of firms with sales in this top range will be used.
3. **Total sales in the last month truncated at the 99th percentile.** As in 2, except truncated at the top 99th percentile.
4. **Total sales in 2012 to date, truncated at the 99th percentile**. BF6 – measured as per 3.
5. **Sales are higher than one year ago.** BF7=3
6. **Total profits in the last month with no truncation:** BF9. For businesses not answering the exact answer, but answer the range question, the midpoint of the range will be used. For firms in the top range, a value equal to the median of firms with sales in this top range will be used.
7. **Total profits in the last month truncated at the 99th percentile.** As in 6, except truncated at the top 99th percentile.
8. **Total profits in the best month of the year, truncated at the 99th percentile.** BF10, measured as per 7.
9. **The inverse hyperbolic sine transformation** of total business profits in the past month $\log(y+(y^2+1)^{1/2})$ – which is similar to the log transformation, but can deal with zero profits. BF9. For businesses not answering the exact answer, but answer the range question, the midpoint of the range will be used. For firms in the top range, a value equal to the median of firms with sales in this top range will be used.
10. **Sales of main product in past month**. BF12b*BF12d
11. **Mark-up profit on main product in past month:** (BF12b-BF12c)*BF12d
12. A standardized **profits and sales** impact will be obtained by aggregating these different effects as described below in our methods section as a standardized z-score.

Example 2:

Business training program

Family A: Business Practices

These questions were asked at both the 1-2 month and 6-8 month follow-ups. However, questions indicated by * were only asked at the longer-term follow-up. For the short-term follow-up measure, the outcome will omit these questions.

- Index of Marketing Practices: the proportion of the following 11 marketing practices currently used by the business:
  - Monitored prices of a competitor's business (*bus4_1_1_6m*=1 *or bus4_1_2_6m=1*)
  - Monitored products of a competitor's business (*bus4_2_1_6m=1 or bus4_2_2_6m=1*)
  - Asked customers if there are other products they would like business to sell (*bus4_3*=1)
  - Spoke with an ex-client to find out why they had stopped purchasing (*bus4_4=1)*
  - Asked a vendor which products sell best (*bus4_5=1)*
  - Used a special offer to attract customers (*bus4_6=1)*
  - Did some form of publicity (*bus4_7=1)*
  - Compared the prices and quality offered by other vendors (*bus4_10=1)*
  - * Performed customer segmentation to help determine marketing strategy (*bus4_11=1).*
  - * Has a logo for their brand (*bus5_17=1)*
  - * Has a registered trademark (*bus5_18=1 or 3 (in progress))*
- Index of accounting and financial practices: the proportion of the following 11 accounting and finance practices used by the business:
  - Keeps written records (*bus5_1=1).* If *bus5* is missing, answer to *bus3* will be used (how do you keep accounts), with answers*<=4* taken to mean written records are kept.
  - Records every purchase and sale (*bus5_2=1)*
  - Uses records to find out how much money business has (*bus5_3=1)*
  - Uses records to know if sales of a product go up or down from one month to another (*bus5_4=1)*
  - Calculates how much each of the major products or services it sells costs the business (*bus5_5=1)*
  - Knows which products/services are the most profitable (*bus5_6=1)*
  - * Pays themselves a salary as an employee of the business (*bus5_9=1)*
  - Have records showing business would have enough money to pay off a loan (*bus5_11=1 bus5_8=1 in short-term follow-up)*
  - Has document detailing annual profits and loss of company (*bus5_14=1, bus5_11=1 in short-term follow-up)*
  - Tracks cash income annually (*bus5_15=1, bus5_12=1 in short-term follow-up)*
  - Separates household and personal finances (*bus5_16=1, bus5_15=1 in short-term follow-up)*

Some practical tips with PAPs

- Don’t overcomplicate and try and specify every eventuality
  - E.g. 1: Some outcomes/mechanisms only make sense/are of interest if you see impact on a key outcome first.
    - For example, a primary outcome of a management improvement program might be the number of workers in the firm.
      - IF you find the number of workers has increased, you might then want to look at whether they are trying new ways of hiring, the wages they pay workers and whether they pay for performance, whether they are hiring young women, etc.
      - IF you find employment falls, you might want to look at whether they are firing workers more versus just reducing hiring, at what types of workers they got rid of, at whether they are now using more capital instead of workers, etc.
      - IF you find no change in employment, then much less interesting to look at all these channels.
    - One approach would be to try and pre-specify complicated IF/ELSE plans, but hard to anticipate everything
    - Instead just focus on main outcomes, and then acknowledge that analysis understanding mechanisms/channels is exploratory.

Some practical tips with PAPs

- *Timing:*
  - Need more information than for AEA registry – so register project "early" (once getting underway) on the registry, and then add PAP after baseline/once understand intervention better (but before any follow-up data). Pilots can help.
  - Can update over time – e.g. add new version before second follow-up.
- *Write-up:* Too mechanical an adherence to the PAP makes for boring papers that include irrelevant information and exclude relevant information
  - Also reason for not specifying too long a PAP – no one wants to read a zillion appendix tables saying we pre-specified we would look at these extra 50 outcomes and 10 types of heterogeneity.
  - I want the authors to bring in descriptive data, qualitative information, new thoughts, and their own exploratory hypotheses to help explain the results they got - just so long as they indicate that this is exploratory and post-hoc
  - Duflo et al. (2020) suggested a separate "populated PAP" that can compliment the paper

Registered reports

- Even more work than a PAP – write up as much of the paper as you can without seeing results, and get it peer-reviewed
- Journal of Development Economics

> Registered Reports: The JDE offers authors the opportunity to have their prospective empirical projects reviewed and approved for publication before the results are known (referred to as 'Registered Reports'). This pre-results review track may be particularly suitable for authors working on research projects for which they have not yet collected or accessed data. Submissions in this track will follow existing policies outlined in the Author Information Pack, including the Mandatory Replication Policy, but specific information is available in the JDE Registered Reports Author Guidelines. A website including the Guidelines and information on Phase 1 acceptances to data is available here ↗ . To submit a Registered Report, select "Registered Report Stage I: Proposal" as the article type in the submission portal. "Registered Report Stage II: Full Article" should only be used for articles derived from accepted Stage I submissions.

Advantages of registered reports

- Get feedback on design and useful/constructive comments at a time where you can do something about it.
- Builds on PAP to get a lot of the write-up done in advance on paper and help you think through carefully what study is doing
- Acceptance without seeing results helps guard against difficulty of publishing null results, and can give you something earlier on to show for study that is going to take years
- May provide some commitment/protection against changes in policy partners objectives/openness to reporting

- *But is a lot of work – like registration/PAPs, makes more sense for longer-term, riskier, hard to replicate projects.*

Some tips on registered reports

- Introduction – should be written like a regular paper, making clear what this paper does, why it is interesting, and how it contributes to what we know from the existing literature
- Reports often lack key specifics:
  - Provide more context and outline the status quo/problem that the intervention is trying to solve: tell us the context and details of the sample, and whether there is a market failure or problem that needs to be solved
    - E.g. if you are doing an experiment on helping youth find jobs- tell us what the status quo process is like, what the background of the youth are, etc.
    - Very common for people to just put in table of summary statistics and not describe the sample at all or offer more details.
  - Unpack the black box – describe in detail the interventions
  - Outline a clear theory of change linked to hypotheses
    - This should also help justify choices of outcomes, and of timing of follow-ups – e.g. is it reasonable to expect impacts in 6 months?
  - Outcomes need to be defined precisely

Some tips on registered reports

- Power calculations often particularly problematic:
  - Discuss assumptions about take-up rates, control means and standard deviations, etc.
  - Discuss effect sizes in meaningful units where possible – e.g. what is the percentage point change in employment you can detect?
  - Relate MDEs to the existing literature to discuss reasonableness – e.g. don't just tell me your MDE is a 8% change in employment – but given the existing literature (and economic theory/cost effectiveness), is it reasonable to think the program will have this effect?

Conclusions

- Very rare for development economists to do study with single treatment and single outcome of interest

(contrast A/B testing in online experiments)

- A variety of different approaches for dealing with multiple testing.
  - “horses-for-courses” – different methods useful for different issues
- Thinking about how you will do this in advance is helpful both for your research, and for credibility
  - Pre-registration, PAP, and RR all helpful here in setting out key outcomes, thinking through what matters most.