

Meta-Analysis

COMM 4940
The Internet



J. Nathan Matias

@natematias

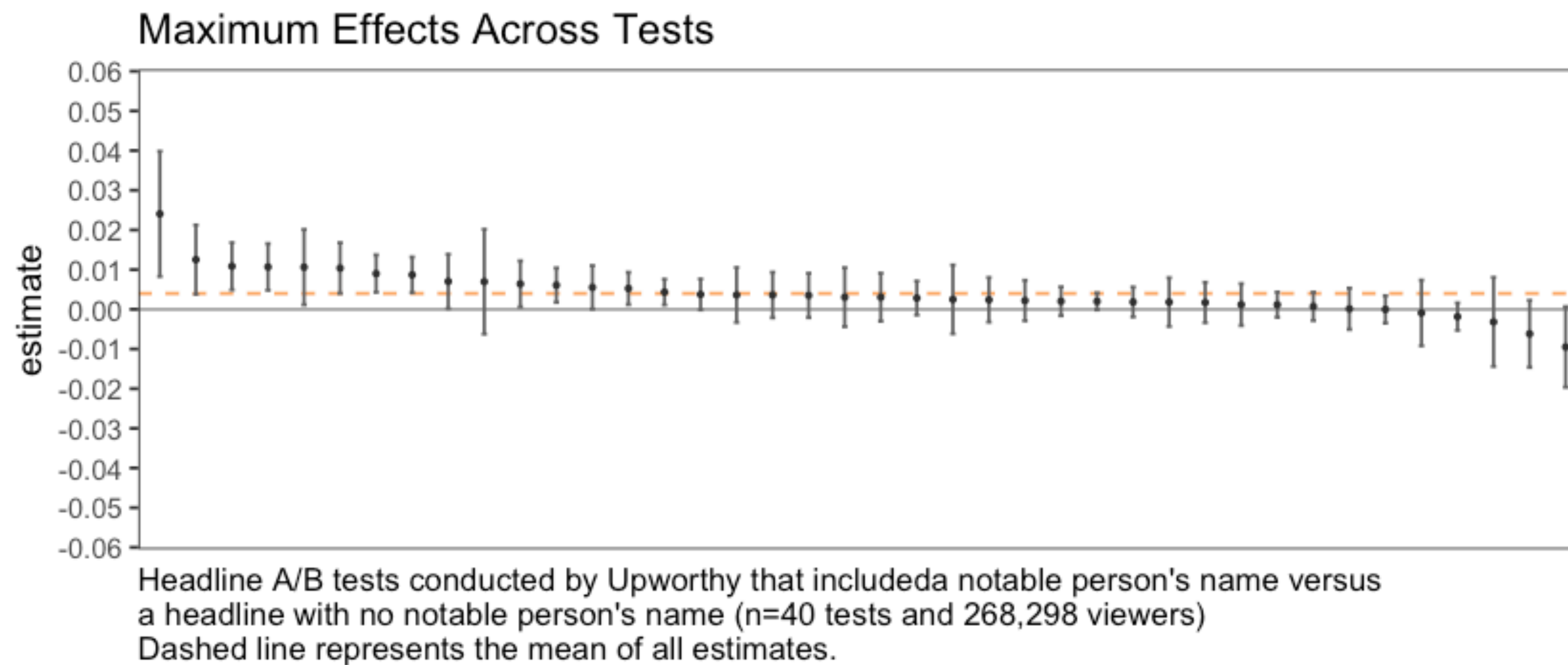
citizensandtech.org

natematias.com

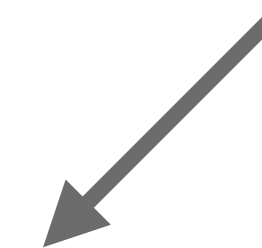
Cornell
CALS

Analyzing Multiple Tests Together

Once we have selected tests, we can analyze the combined effect across many tests. The practice of combining studies is called “meta-analysis.”

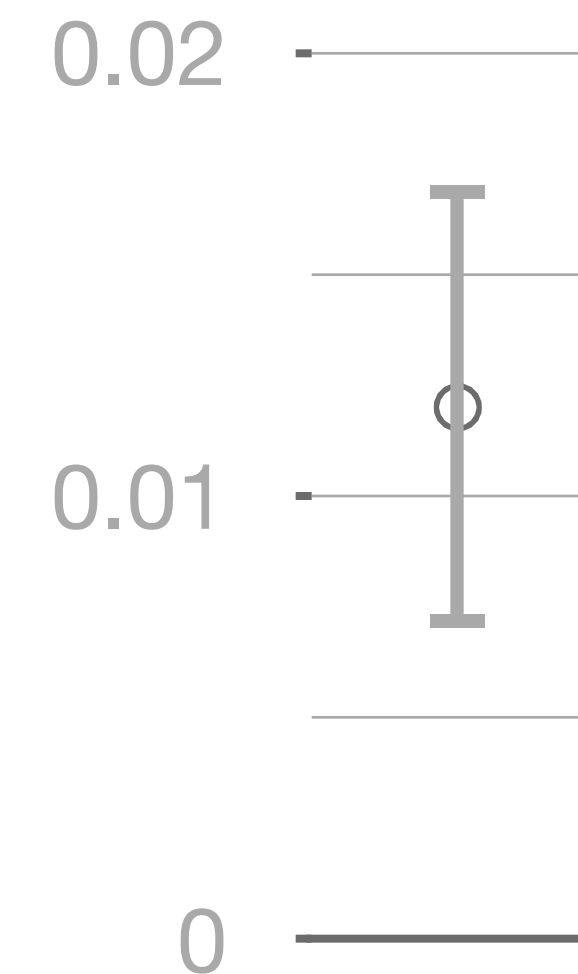
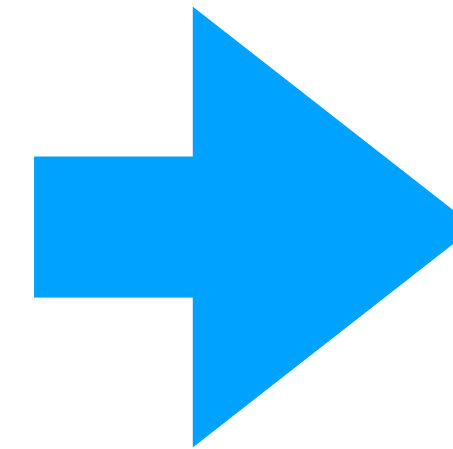
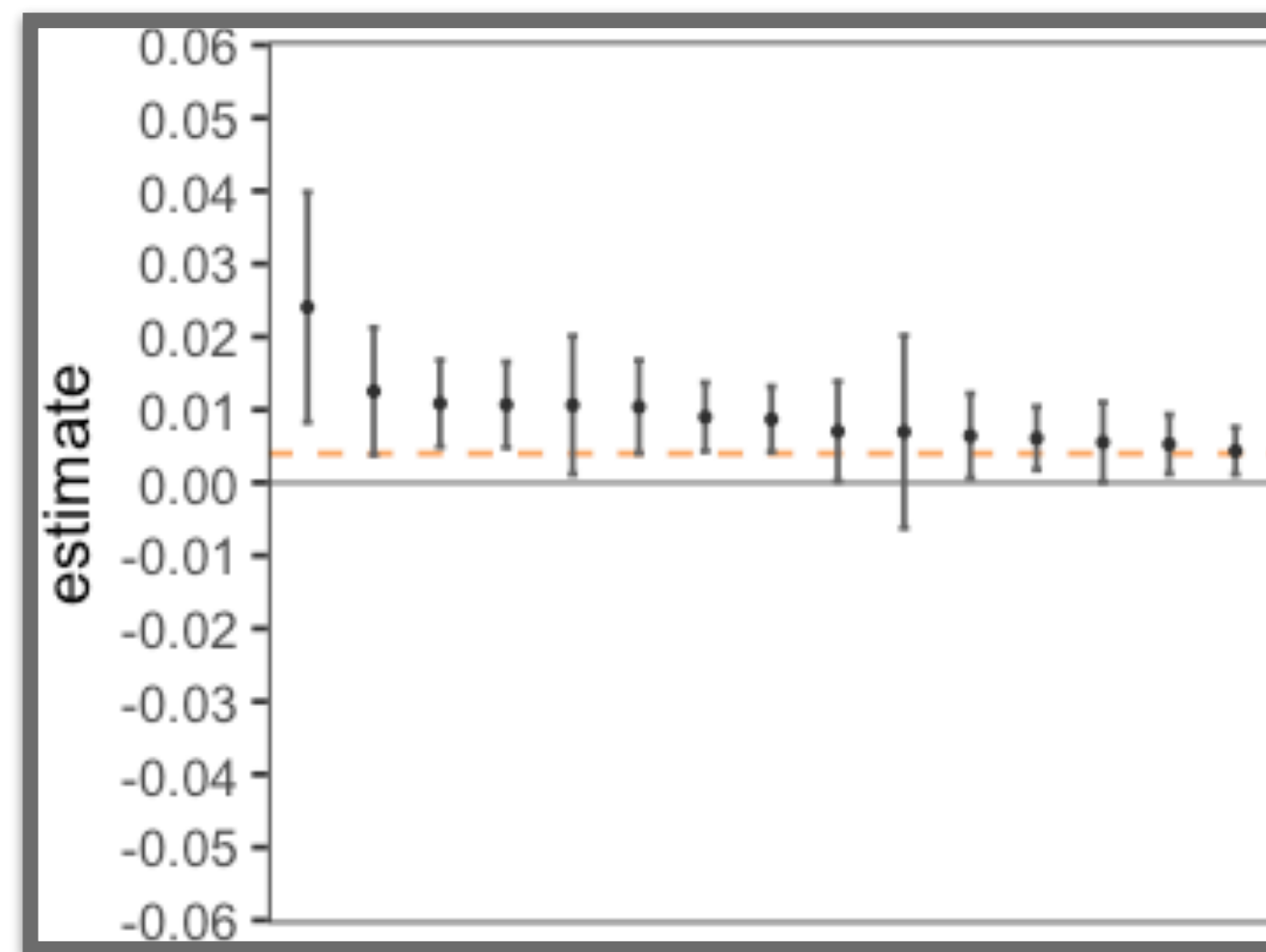


Each of these
is the effect size
of including a
notable person's name
in the headline

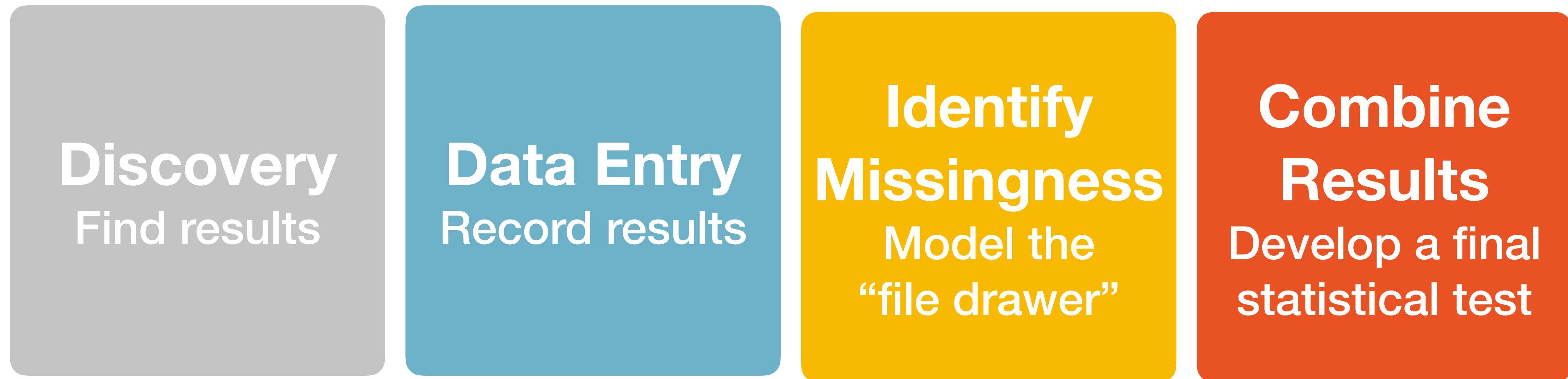


What is Meta-Analysis?

Meta-analysis combines the results of many studies into one result, or into a pattern of results.



What is Meta-Analysis?



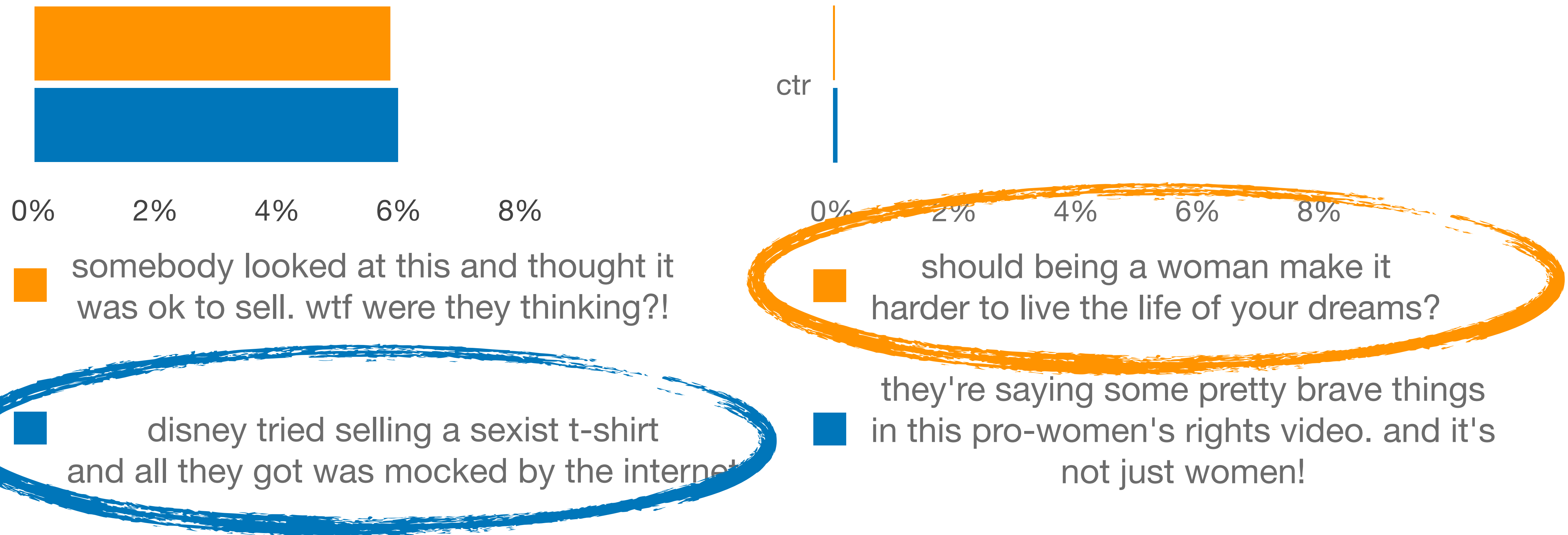
**With the Upworthy Archive
We get to focus on the last step!**



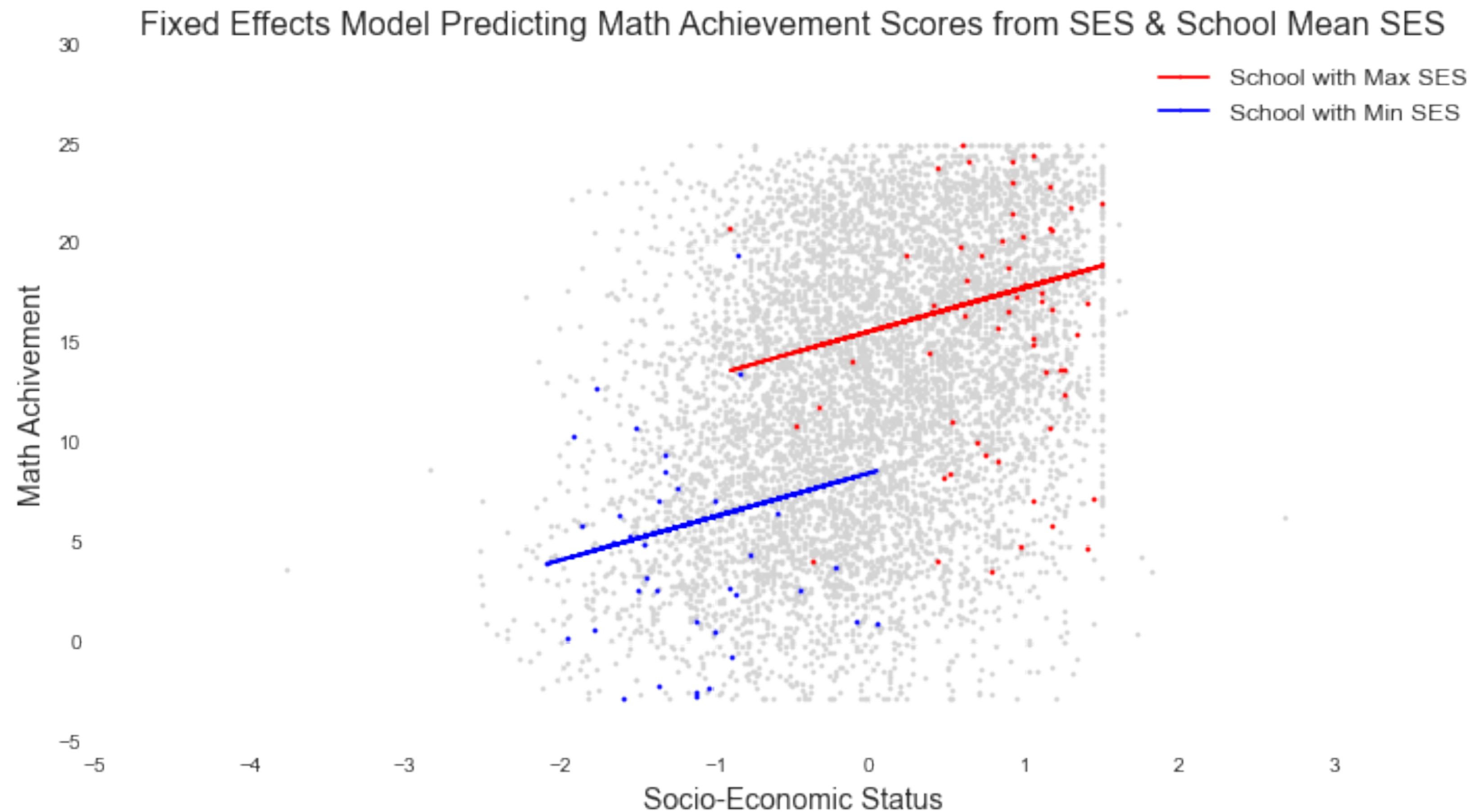
Why We Can't Just Merge the Data Do a Statistical Test, and be Done

Example hypothesis: **phrasing a headline as a question** increases the chance that an article will be **clicked on**

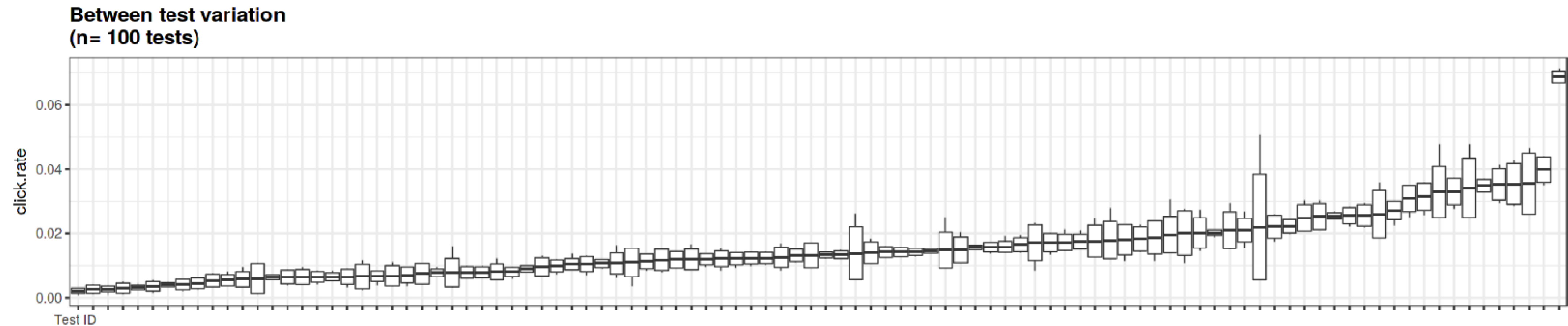
Comparing Within Tests Vs Comparing Between Tests



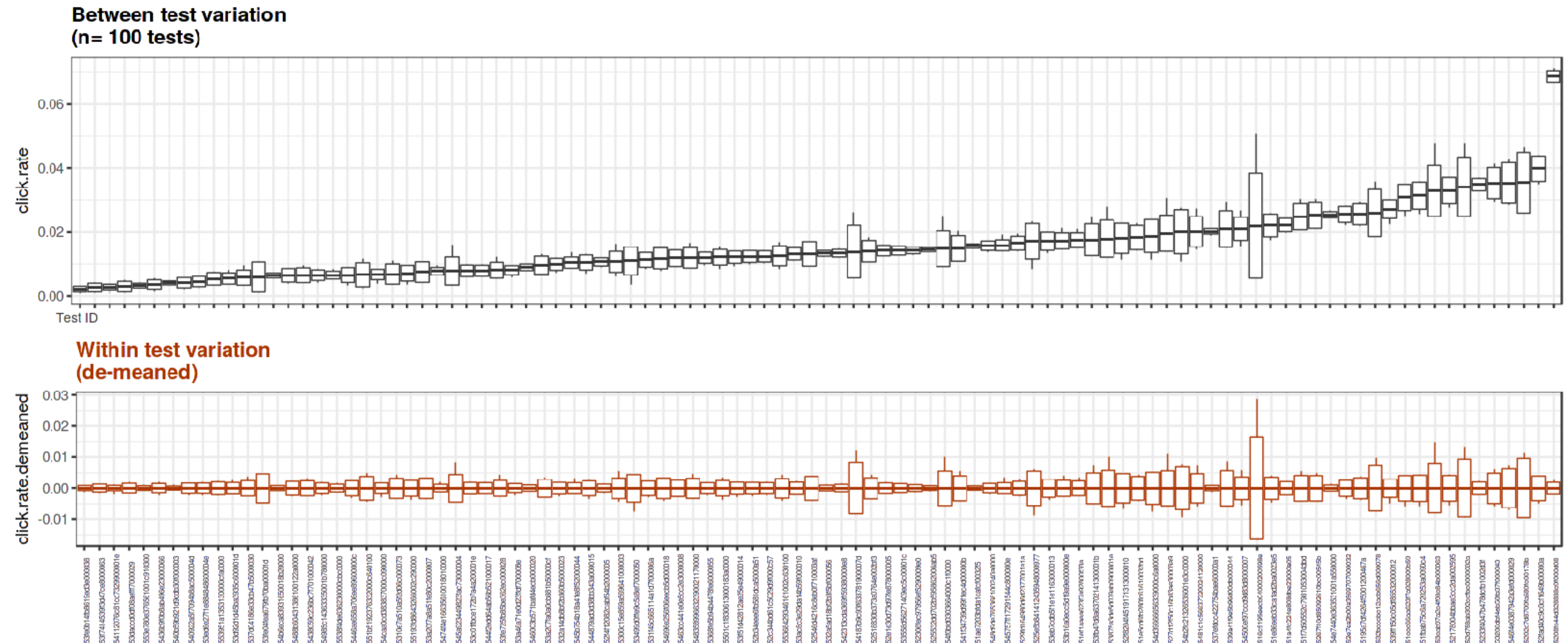
Within-Group Variation & Between-Group Variation



Comparing Within Tests Vs Comparing Between Tests



Comparing Within Tests vs Comparing Between Tests



Fixed Effects Model

Important: “Effects” in “Fixed effects” has nothing to do with causality. Instead, it refers to the variation between groups.

Linear Regression Model

The diagram illustrates the components of the linear regression equation $y = \beta_0 + \beta_1 X + \epsilon$. The equation is written in a large, dark grey font. Above the equation, four labels in red text are positioned: "Outcome" on the left, "Intercept" above β_0 , "Predictor" above $\beta_1 X$, and "Error / Residual" above ϵ . Red arrows point from each label to its corresponding term in the equation: from "Outcome" to y , from "Intercept" to β_0 , from "Predictor" to $\beta_1 X$, and from "Error / Residual" to ϵ .

Outcome

Intercept

Predictor

Error / Residual

$$y = \beta_0 + \beta_1 X + \epsilon$$

Fixed Effects Model

The diagram illustrates the Fixed Effects Model equation: $y_{ij} = \beta_0 + \beta_1 X_{ij} + \mathbf{u}_j + \epsilon_{ij}$. Red arrows point from labels to the corresponding terms in the equation. 'Outcome' points to y_{ij} . 'Intercept' points to β_0 . 'Predictor' points to X_{ij} . 'Group Error / Residual' points to \mathbf{u}_j . 'Individual Error / Residual' points to ϵ_{ij} . Additionally, 'Individual' and 'Group' arrows point to X_{ij} , indicating its level of measurement.

Outcome

Intercept

Predictor

Group Error / Residual

Individual Error / Residual

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + \mathbf{u}_j + \epsilon_{ij}$$

Individual

Group

Important: Fixed Effects models make no assumptions about the distribution of group-level means and cannot test hypotheses about group differences

Example hypothesis: **phrasing a headline as a question** increases the chance that an article will be **clicked on**

Testing our Hypothesis with Linear Regression

```
lm(formula = click.rate ~ question, data = packages.subsample.df)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.014274	-0.007341	-0.002041	0.004883	0.056163

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.015188	0.000541	28.073	<2e-16 ***
question	-0.001852	0.001350	-1.372	0.171

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01071 on 465 degrees of freedom
Multiple R-squared: 0.00403, Adjusted R-squared: 0.001888
F-statistic: 1.882 on 1 and 465 DF, p-value: 0.1708

Adding a Variable for Every Test

```
lm(formula = click.rate ~ question + clickability_test_id, data = packages.subsample.df)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.010504	-0.001804	-0.000125	0.001627	0.028066

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	0.0226497	0.0017968	12.605
question	-0.0017017	0.0008068	-2.109
clickability_test_id5195a7df4264f5001200467a	0.0029387	0.0026710	1.100
clickability_test_id51a4fc224e808dbe23000a05	-0.0004285	0.0026710	-0.160
clickability_test_id51ae7203bfdalcafcf0032f5	-0.0070362	0.0033226	-2.118
clickability_test_id51e5dd6fb280fc61610032bd	-0.0039678	0.0026542	-1.495
clickability_test_id51e86eeb33ca1ad2ba0023e5	-0.0001475	0.0026542	-0.056
clickability_test_id51eec66aa623f7a928009e59	0.0082122	0.0026710	3.075
clickability_test_id51ef1aaea623f7a92800f33a	-0.0035074	0.0026952	-1.301
clickability_test_id51f7d50552c7981053004dbd	0.0038781	0.0026952	1.439
clickability_test_id51fbab75c6a720253a0090c4	0.0087741	0.0026710	3.285
clickability_test_id5202c7d8709fa589fe00178b	0.0128202	0.0026710	4.800
clickability_test_id5208fe44d328838e9f000d08	0.0460728	0.0026710	17.249
clickability_test_id52178004bbae3cc2da002595	0.0103842	0.0026710	3.888
clickability_test_id52306fec97956ef5290009e0	-0.0066895	0.0026952	-2.482
clickability_test_id5233f4047b478dcfb1002d3f	0.0122009	0.0026710	4.568

Think of these
coefficients as u_j

Adding a Variable for Every Test

```
lm(formula = click.rate ~ question + clickability_test_id, data = packages.subsample.df)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.022649721	0.0017968176	12.605465	1.659855e-30
question	-0.001701737	0.0008067755	-2.109307	3.559677e-02

The result is now
statistically-significant
(p=0.035)

Specifying a Fixed Effects Model

```
plm(formula = click.rate ~ question, data = packages.subsample.df,  
     model = "within", index = "clickability_test_id")
```

Unbalanced Panel: n = 100, T = 2-11, N = 467

Residuals:

Min.	1st Qu.	Median	3rd Qu.	Max.
-0.01050412	-0.00180374	-0.00012501	0.00162732	0.02806549

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
question	-0.00170174	0.00080678	-2.1093	0.0356 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 0.0057872

Residual Sum of Squares: 0.0057177

R-Squared: 0.01201

Adj. R-Squared: -0.25793

F-statistic: 4.44917 on 1 and 366 DF, p-value: 0.035597

of tests

per group

U_j

of headlines

The result is identical
to the linear regression
with categorical variables
($p=0.035$)

Full Meta-Analysis Example

Example hypothesis: **including a notable person's name in a headline** increases the chance that an article will be **clicked on**

Dataset: **40** tests. **80** headlines. **270,177** viewers.

Linear Model

```
lm(formula = clicked ~ has_treatment, data = participant.df)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.01865	-0.01865	-0.01497	-0.01497	0.98503

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.0149664	0.0003505	42.704	< 2e-16	***
has_treatment	0.0036828	0.0004947	7.445	9.75e-14	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1286 on 270175 degrees of freedom

Multiple R-squared: 0.0002051, Adjusted R-squared: 0.0002014

F-statistic: 55.42 on 1 and 270175 DF, p-value: 9.746e-14

Fixed Effects Model

```
plm(formula = clicked ~ has_treatment, data = participant.df,  
     model = "within", index = "clickability_test_id")
```

Unbalanced Panel: n = 40, T = 2025-20505, N = 270177

Residuals:

Min.	1st Qu.	Median	3rd Qu.	Max.
-0.0708440	-0.0182503	-0.0121964	-0.0074033	0.9976035

Coefficients:

	Estimate	Std. Error	t-value	Pr(> t)
has_treatment	0.0033261	0.0004914	6.7685	1.304e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 4404.5

Residual Sum of Squares: 4403.8

R-Squared: 0.00016956

Adj. R-Squared: 2.1512e-05

F-statistic: 45.8122 on 1 and 270136 DF, p-value: 1.3042e-11