“Hunting out the real uncertainty*”

Yoav Benjamini

Tel Aviv University

‘On the Foundations of Applied Statistics’

Dedicated to the memory of Edna Schechtman

April 2024

*A chapter in Mosteler & Tukey’s ‘Data Analysis and Regression’ book

Tukey's last published work

A puzzling entry on Multiple Comparisons

for the International Encyclopedia of Statistics in the Social Sciences.

with Jones & Lewis, post-mortem 2002

Some general statement about MCP importance

1. FDR in pairwise comparisons

Williams Jones & Tukey '99

2. Analysis of Variance

"Two alternatives, 'fixed' and 'variable', are not enough. A good way to provide a reasonable amount of realism is to define '**c**' by

appropriate error term = f-error term
+**c** [r-error term - f-error term]

It pays then to learn as much as possible about values of c in the real world"

The idea stems from sampling without replacement

From a finite population (Cornfield & Tukey 1956)

What's that to do with multiple comparisons?

YB

Replicability in Genes & Behaviour

Crabbe et a (Science, '99) compared 12 measures across strains at 3 labs

In spite of strict standardization,

Significant Lab*Genotype Interaction

"Thus, experiments characterizing mutants may yield results that are idiosyncratic to a particular laboratory."

Will our computational tools solve the problem?

Comparing 17 measures between 8 inbred strains of mice

At 3 labs: Golani at TAU, Elmer MPRC, Kafkafi NIDA[1]

Kafkafi, YB, Sakov, Elmer, Golani PNAS '04

Significance of 8 Strain differences

Behavioral Endpoint	Labs Fixed
Prop. Lingering Time	0.00001
# Progression segments	0.00001
Median Turn Radius (scaled)	0.00001
Time away from wall	0.00001
Distance traveled	0.00001
Acceleration	0.00001
# Excursions	0.00001
Time to half max speed	0.00001
Max speed wall segments	0.00001
Median Turn rate	0.00001
Spatial spread	0.00001
Lingering mean speed	0.00001
Homebase occupancy	0.001
# stops per excursion	0.0028
Stop diversity	0.027
Length of progression segments	0.44
Activity decrease	0.67

Strain x Lab Interaction significant

FDR ≤ .05

Strain x Lab Interaction not significant

The model and Mixed ANOVA

$Y_{gli} = \mu_g + a_l + b_{gl} + \varepsilon_{gli}$ g=1,…,G; l=1,…,L i=1,…n

Source	df	MSE	F	p-value
Strain	7	102.5	14.8	0.0028
Lab	2	6.35	0.9	0.43
Lab*Strain	14	6.87	3.00	0.00028
Residuals	264	2.29		

The threshold for significant strain differences can be much higher

YB

Recalling Mann's warning in "Behavior Genetics in transition" (Science, 94)

"…jumping too soon to discoveries.." (and press discoveries)
"raises the issue of *Replicability"*

The Encyclopedia's entry is about replicability of discoveries

Addressing : Selective Inference
The relevant variability

Tukey's two statistical pillars of replicability

But be ready to compromise

Addressing the relevant variability

The model and Mixed ANOVA

$Y_{gli} = \mu_g + a_l + b_{gl} + \varepsilon_{gli}$ g=1,…,G; l=1,…,L i=1,…n

Source	df	MSE	F	p-value
Strain	7	102.5	14.8	0.0028
Lab	2	6.35	0.9	0.43
Lab*Strain	14	6.87	3.00	0.00028
Residuals	264	2.29		

The threshold for significant strain differences can be much higher

Esimate of $\sigma^2_{LAB*STRAIN}$ *and of* $\sigma^2_{LAB*STRAIN} / \sigma^2$

YB

II. Addressing the relevant variability

Mouse phenotyping example: opposite single lab results

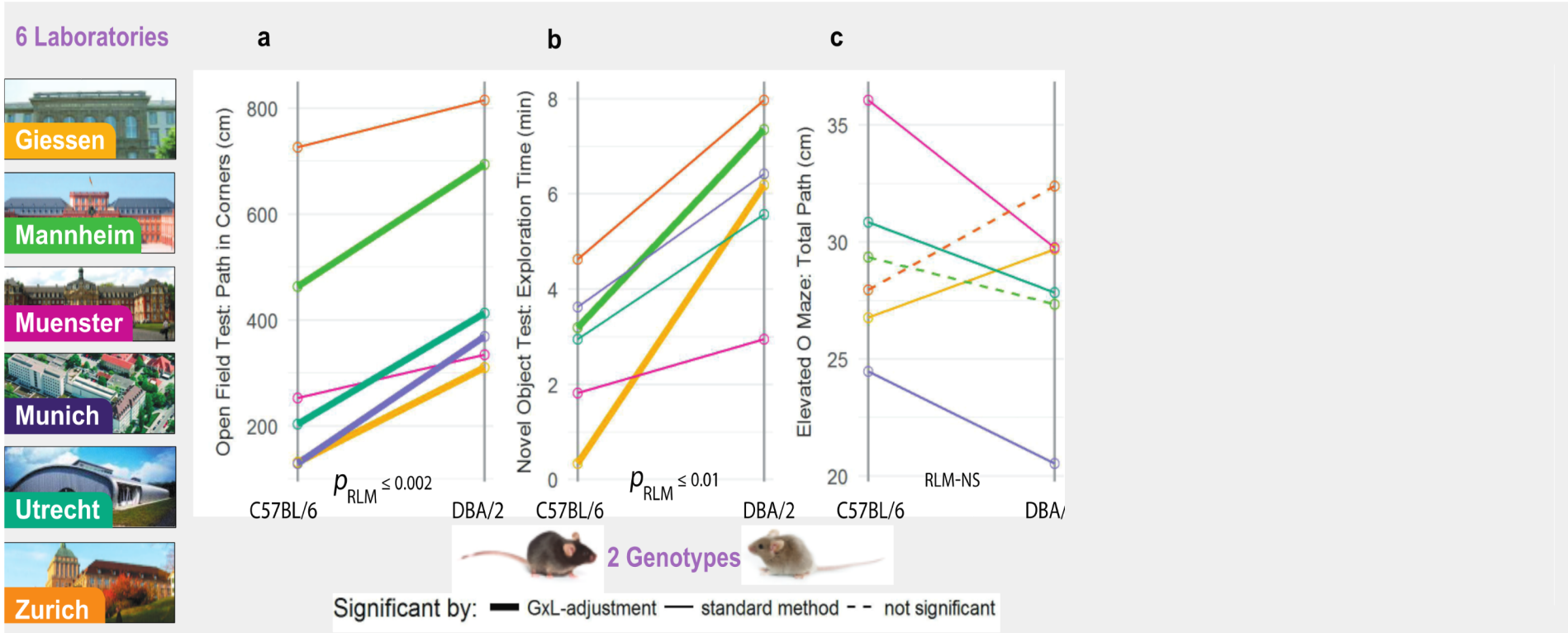


Figure 1 | Genotype-by-Laboratory interaction (G×L). Comparing 2 genotypes across 6 laboratories (coded by color), using three phenotypes out of dataset 1 **(Supplementary Table 1).** Each line connects genotype means within the same laboratory, so its slope reflects their difference. Dashed/ thin lines denote within-lab non-significance/significance using the standard t-test. Bold lines denote significance after GxL-adjustment (all at 0.05). **a.** illustrates significant genotype effect according to the Random Lab Model (RLM) with similar slopes indicating a small GxL effect. **b** illustrates more variation of the laboratory lines, yet the genotype effect appears fairly replicable, and is significant according to the RLM. **c** exhibits substantial GxL: using the standard single-lab analysis Giessen would have reported DBA/2 significantly larger than C57BL/6, while Mannheim, Muenste and Munich would have reported the opposite significant discovery. Such "opposite significant" **(Supplementary Methods S1.1.3)** cases were not rare using the standard method, but disappeared after GxL-adjustment. **d. GxL-adjustment decreases non-replicable discoveries in 8 multi-lab datasets:** average single-lab Type-I error rate .using the standard t-test is much higher than the prescribed 5%. The GxL-adjustment brings it close to 5%, **see Supplementary Table 1**

Kafkafi et al ('17 Nature Methods)

GxL interaction is “a fact of life”

Genotype-by-Lab effect for a genotype in a new lab is not known; but If its variability σ^2_{GxL} can be estimated, use

$$\frac{Mean(Yg_1l_1.) - Mean(Yg_2l_1.)}{(\sigma^2_{Within}\ (1/n+1/n) + 2\sigma^2_{GxL}\ )^{1/2}}$$

We call it GxL- adjustment

It’s the right “yardstick” against which genetic differences should be compared, when concerned with replicability.

Single-lab analyses in all known replication studies

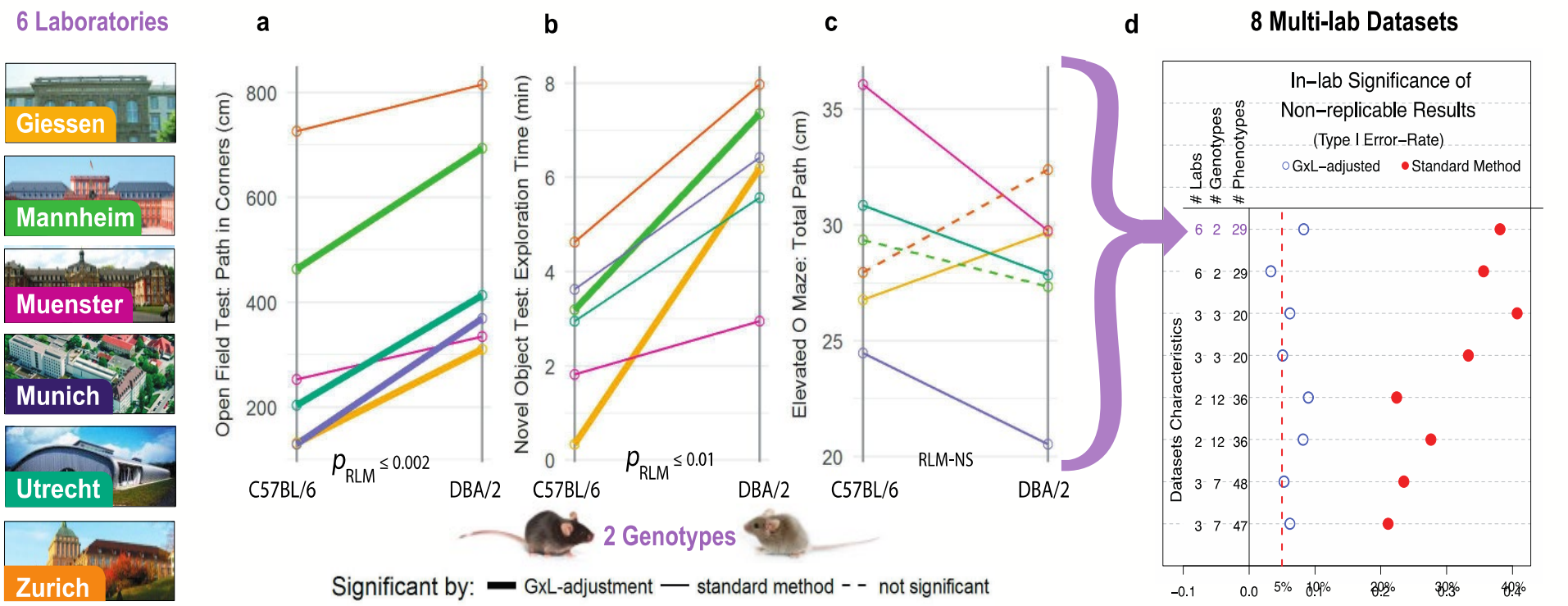


Figure 1 | Genotype-by-Laboratory interaction (G×L). Comparing 2 genotypes across 6 laboratories (coded by color), using three phenotypes out of dataset 1 **(Supplementary Table 1)**. Each line connects genotype means within the same laboratory, so its slope reflects their difference. Dashed/ thin lines denote within-lab non-significance/significance using the standard t-test. Bold lines denote significance after GxL-adjustment (all at 0.05). **a.** illustrates significant genotype effect according to the Random Lab Model (RLM), with similar slopes indicating a small G×L effect. **b** illustrates more variation of the laboratory lines, yet the genotype effect appears fairly replicable, and is significant according to the RLM. **c** exhibits substantial G×L: using the standard single-lab analysis Giessen would have reported DBA/2 significantly larger than C57BL/6, while Mannheim, Muenster and Munich would have reported the opposite significant discovery. Such "opposite significant" **(Supplementary Methods S1.1.3)** cases were not rare using the standard method, but disappeared after GxL-adjustment. **d. GxL-adjustment decreases non-replicable discoveries in 8 multi-lab datasets:** average single-lab Type-I error rate **.**using the standard t-test is much higher than the prescribed 5%. The GxL-adjustment brings it close to 5%, **see Supplementary Table 1**

Kafkafi et al ('17 Nature Methods)

Utilizing large database

Extract the relevant the GxL-factor γ per endpoint from a public database

$$\gamma = \sigma_{GxL} / \sigma_{within}$$

"Replicability Adjuster" Implemented at the Mouse Phenotyping Database (MPD) in JAX Bar Harbor

Kafkafi et al (Nature Methods '17)

YB

"Two alternatives, 'fixed' and 'variable', are not enough.
A good way to provide a reasonable amount of realism
is to define '**c**' by

appropriate error term = f-error term
+**c** [r-error term - f-error term]

It pays then to learn as much as possible
about values of c in the real world"

In none of our work could we have a random sample of labs

Still

Better treat as 'random' than as fixed

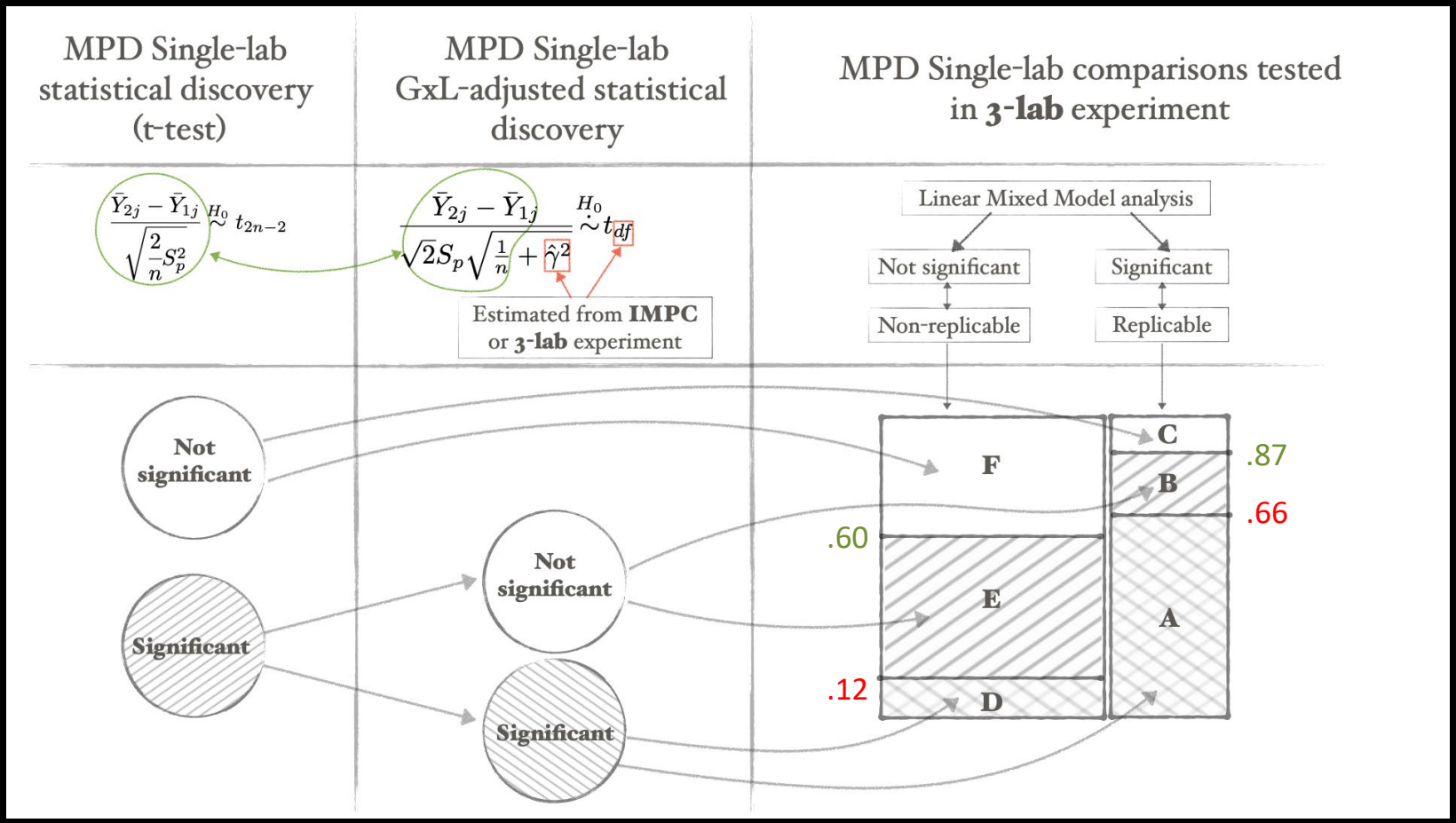
YB

Testing the approach

- Took 165 Single lab experimental results involving comparisons between mouse strains from Mouse Phenotyping Database
- Carried similar experiments in 3 labs : JAX, TAUL, and TAUM without much coordination
- Used Random (variable) Lab Mixed Model Analysis to assess replicability of original results

Estimated $\gamma^2 = \sigma^2_{GxL} / \sigma^2_{within}$ from IMPC or from our experiments

Compared original results with their GxL adjusted results



Type I replicability error: Original analyses 60% GxL adjusted 12%

Reducing to 0.005 24%

Jaljuli, Kafkafi et al ’23

Research supported by ISF- NIH grants

The GxL Factor γ^2 is per endpoint

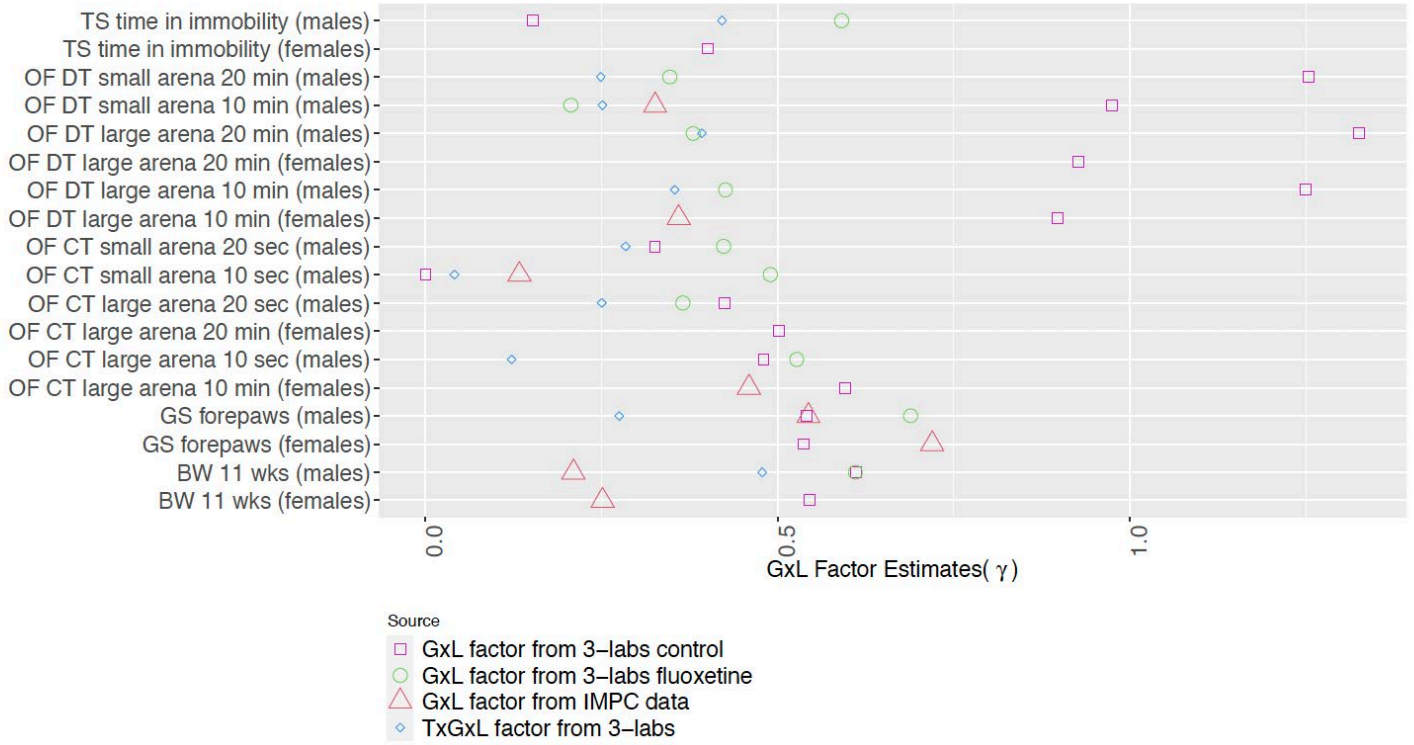
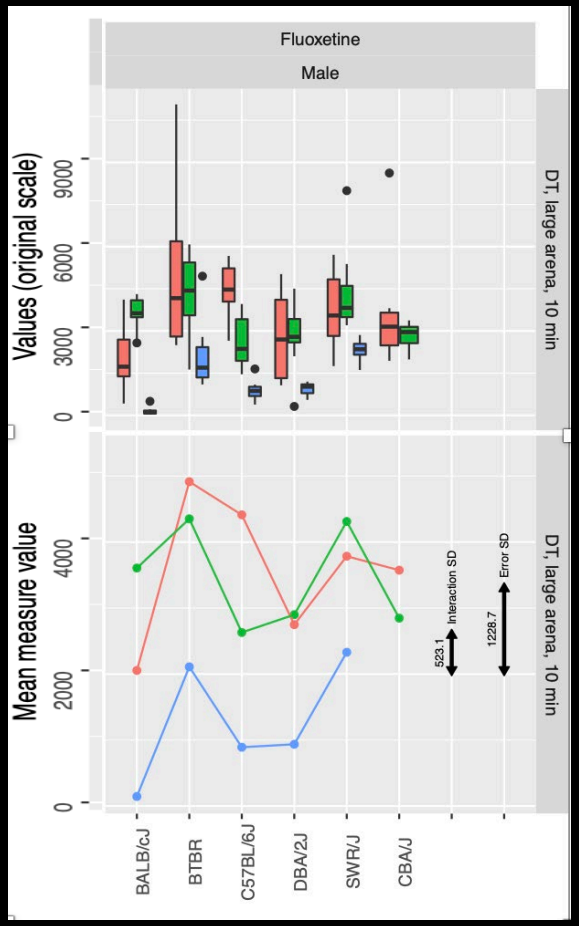


Figure 3: The values of the estimated interaction factor γ, for all endpoints as estimated from various sources: GxL factor from our 3-labs control data and from our fluoxetine treated data; GxL factor from IMPC data; TxGxL factor from our 3-labs data. CT and TS were logit transformed and GS was raised to the power of 1/3.

The GxL Factors are large!

We also experimented with drugs



Improving measures by
Reducing the $\gamma^2 = \sigma^2_{GxL} / \sigma^2_{within}$

Many small studies are better than a few large ones

That's what we have in Meta-analysis of Systematic Reviews

Taking the lesson to meta-analysis

Common effect analysis vs Random effect analysis

Fixed ANOVA Mixed

Decision based on measures of between study variability.

Our Lesson: Use always Random Effects

But compromise (per Tukey's advice)

Indeed, Gaussian dist'n is used rather than t_{df} df very small

For animal studies Gaussian assumption (after transformation) is reasonable. For clinical trials?

Sometimes Yes; Sometimes No (then use Jaljuli, …, Heller et al '22)

Taking the lesson to cross validation /Jackknife/ Data Splitting

Do not always divide by random sampling of cases/observations

Identify the source of variation relevant to the user:

Year-to-year; Institutions; Locations; People

Divide the groups in k-fold cross validation accordingly

Mosteller & Tukey emphasized this point for Jackknife estimator

Camil Fuchs, when developing election night prediction

With only 3 elections data available,

Developed Model on 1&2 tested on 3

" 2&3 " 1

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Developing a length of stay prediction model for newborns, achieving better accuracy with greater usability

Tzviel Frostig [a,*], Yoav Benjamini [a,b], Orli Kehat [c], Ahuva Weiss-Meilik [c], Dror Mandel [d], Ben Peleg [e,f], Zipora Strauss [e,f], Alexis Mitelpunkt [e,g]

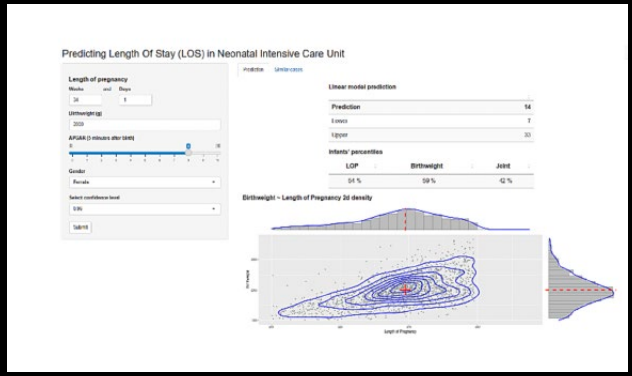
- Data from TAMC sorted by year

  Training set

  2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018

  Test set

- External validation set from Sheba Medical Centre



Take away massages

Replicability can be enhanced mainly by addressing

Selective inference

- The silent killer of replicability (YB ‘22 HDSR)

The relevant variability

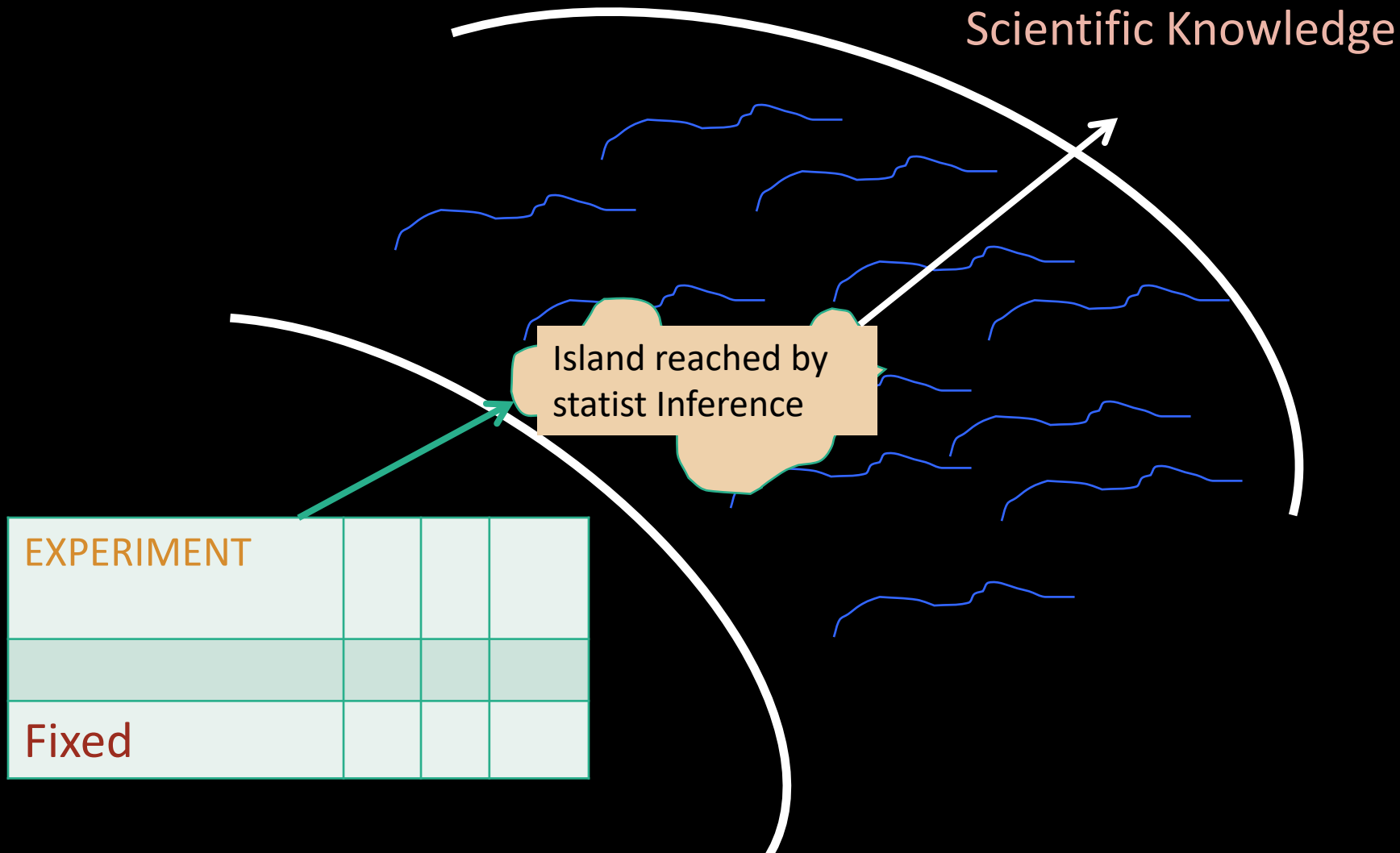
- Identify the relevant sources of variability
- Prefer random model analysis even if levels are only ‘variable’
- Do not shy away from out of study estimates
- Many small studies are better than a single large one

Do not give up addressing both, but do not be afraid to compromise

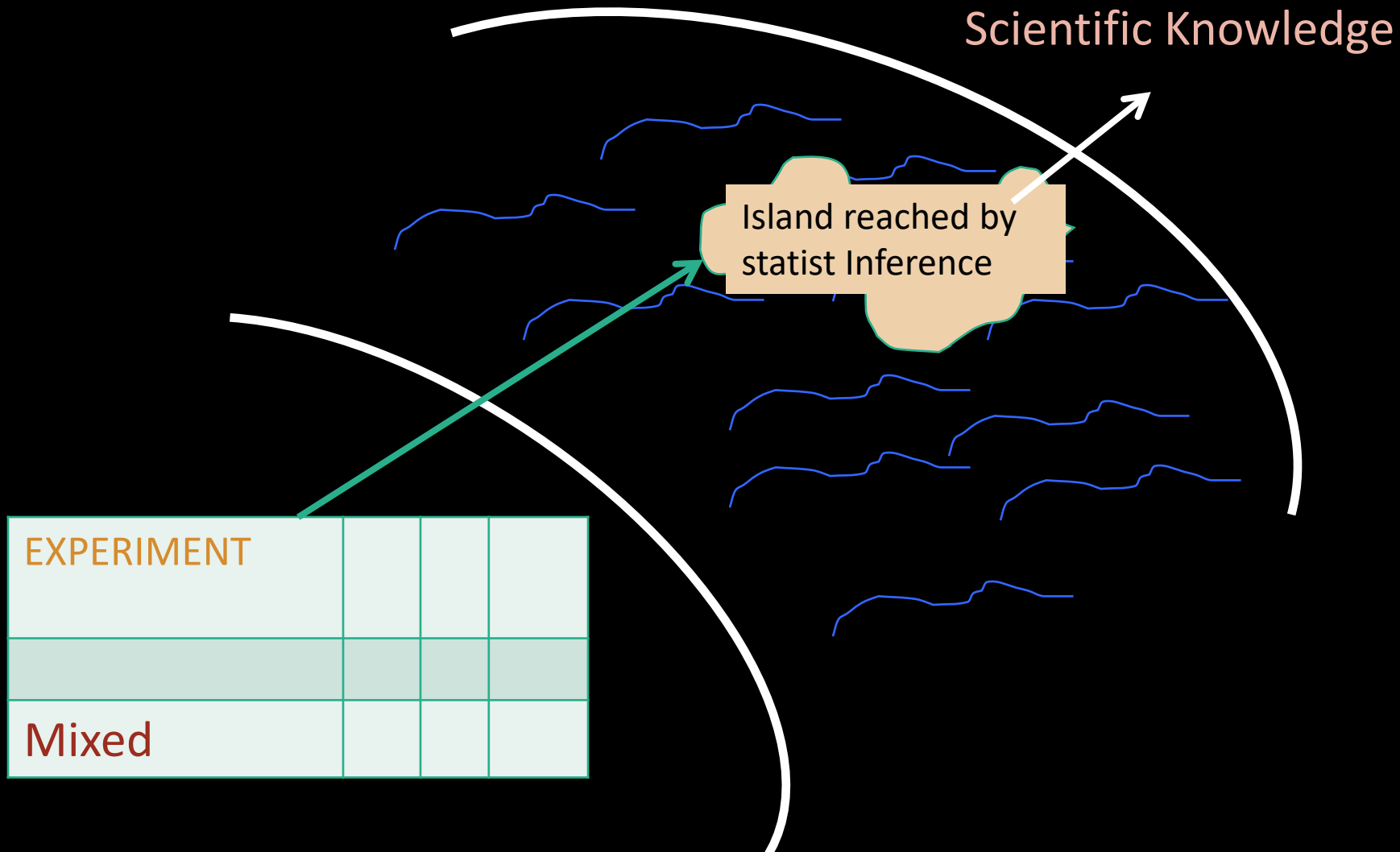
- Secondary endpoints in clinical and epidemiological studies

Thanks to JWT for the insight

Reading ‘56 paper again



Reading ‘56 paper again



In Memory of

Edna Schechtman

1948-2022

Ilan Golani

1935-2024

Thanks

1888 1999

The industrialization of the scientific process

1950 2010