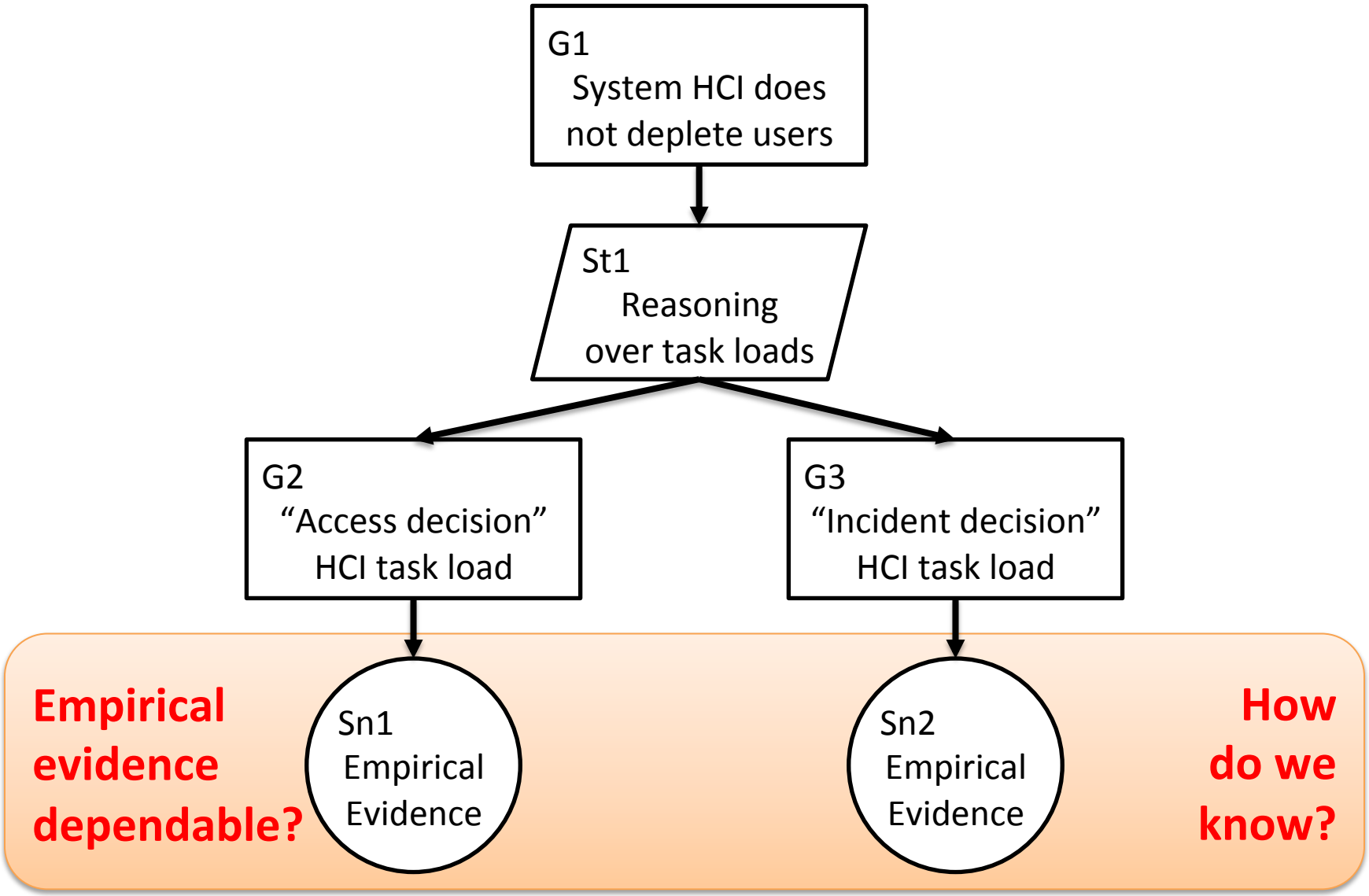
Dependable Research on the Human Dimensions of Cyber Security

71st IFIP WG10.4 Meeting

Thomas Gross
Newcastle University

Motivation
Assurance Case Considering Humans

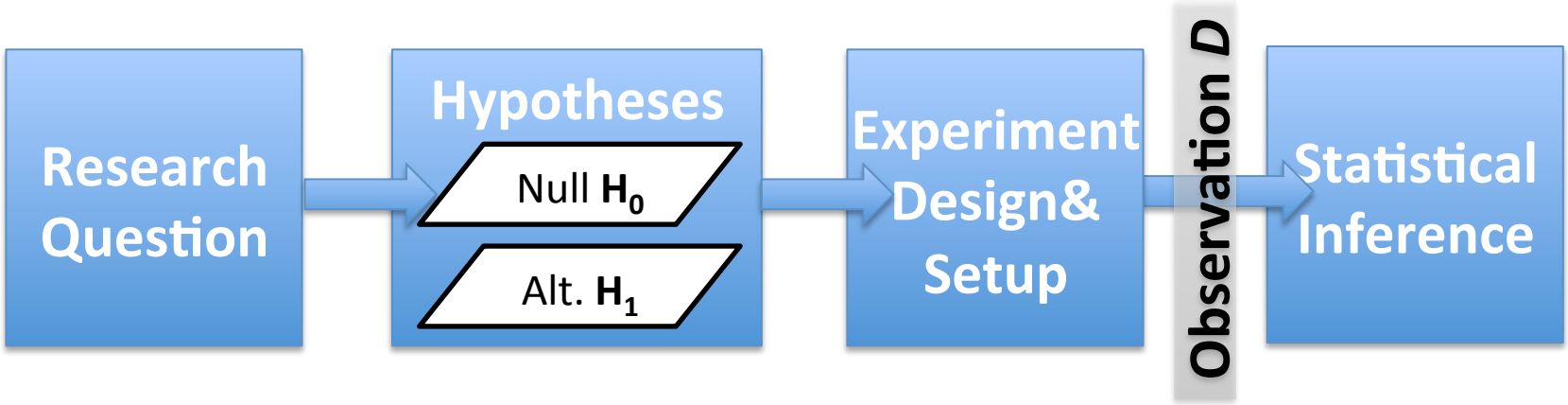


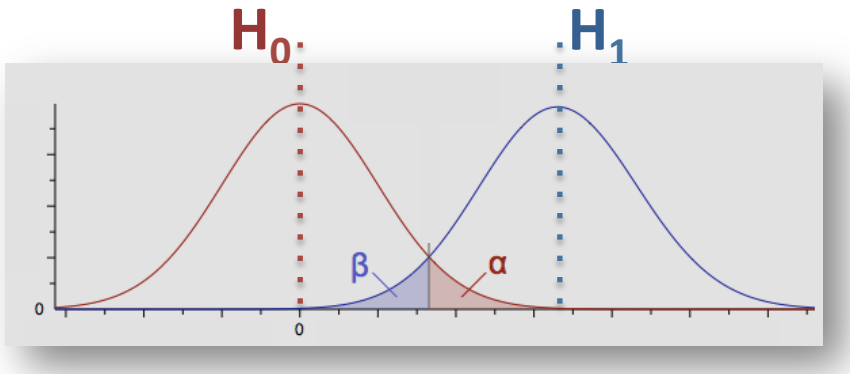
Dependable Evidence on Human Dimensions of Cyber Security

- How can we ascertain the dependability of evidence?
- What research methods need to be employed to yield dependable evidence?
- How can we truly advance knowledge on human dimensions of cyber security?

SETTING THE STAGE

Statistical Hypothesis Testing





[Lehmann and Romano (2005) Testing Statistical Hypotheses]

Gain Knowledge from Experiments

Given the observation *D* of an experiment, infer the likelihood of the null hypothesis $\mathbf{H_0}$.

Pr($\mathbf{H_0}$ | *Observation D*)?

p-Values

***p*-Value:** Likelihood of observing a result equal or more extreme as the observation *D*, assuming the null hypothesis $\mathbf{H_0}$.

$$p = \mathbf{Pr(}\text{Observation } D \mid \text{Null Hypothesis } \mathbf{H_0}\mathbf{)}$$

Statistical Significance:
if $p <$ pre-defined significance level α ➔ reject $\mathbf{H_0}$.

… and their pitfalls

p-Value says little about the actual likelihood of the null hypothesis after empirical observation *D*.

$$p = \mathbf{Pr(}\text{Observation } D \mid \mathbf{H_0)}$$

$$\neq \mathbf{Pr(H_0} \mid \text{Observation } D\mathbf{)}$$

Pitfalls: Often misinterpreted or misreported.

Good practice: Effect sizes & confidence intervals.

[cf. Nickerson – Null Hypothesis Significance Testing – A Review of an Old and Continuing Controversy]

Positive Predictive Value

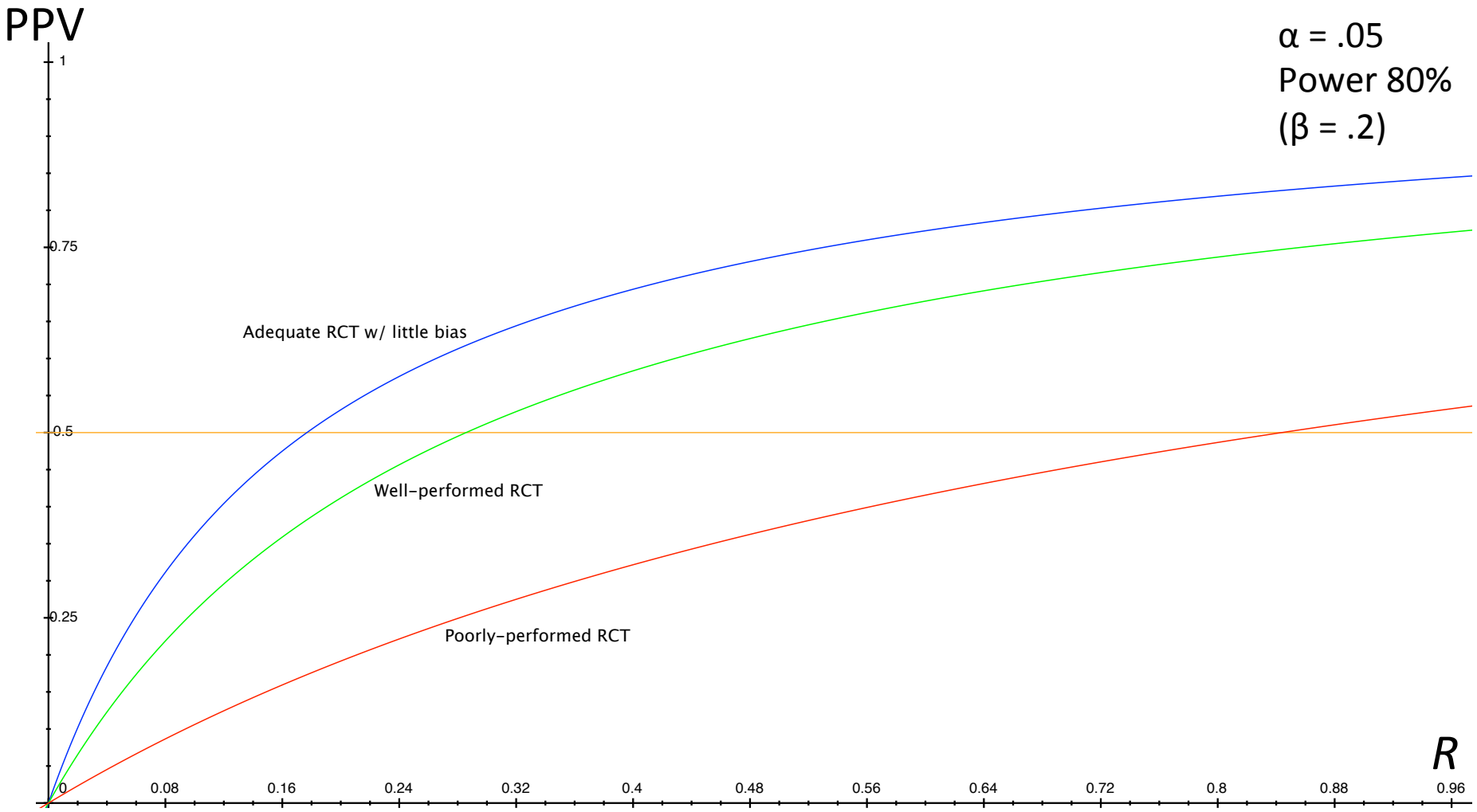
Positive Predictive Value (PPV): The likelihood of the alternative hypothesis $\mathbf{H_1}$ being true in reality, given the experiment's observation *D*.

Derived with Bayes Theorem. Impacted by:

- Significance level α
- Power 1-β
- Prior likelihood of alternative hypothesis *R*=$\mathbf{Pr(H_1)}$
- Bias in the experiment *u* (Estimate cf. Ioannidis)

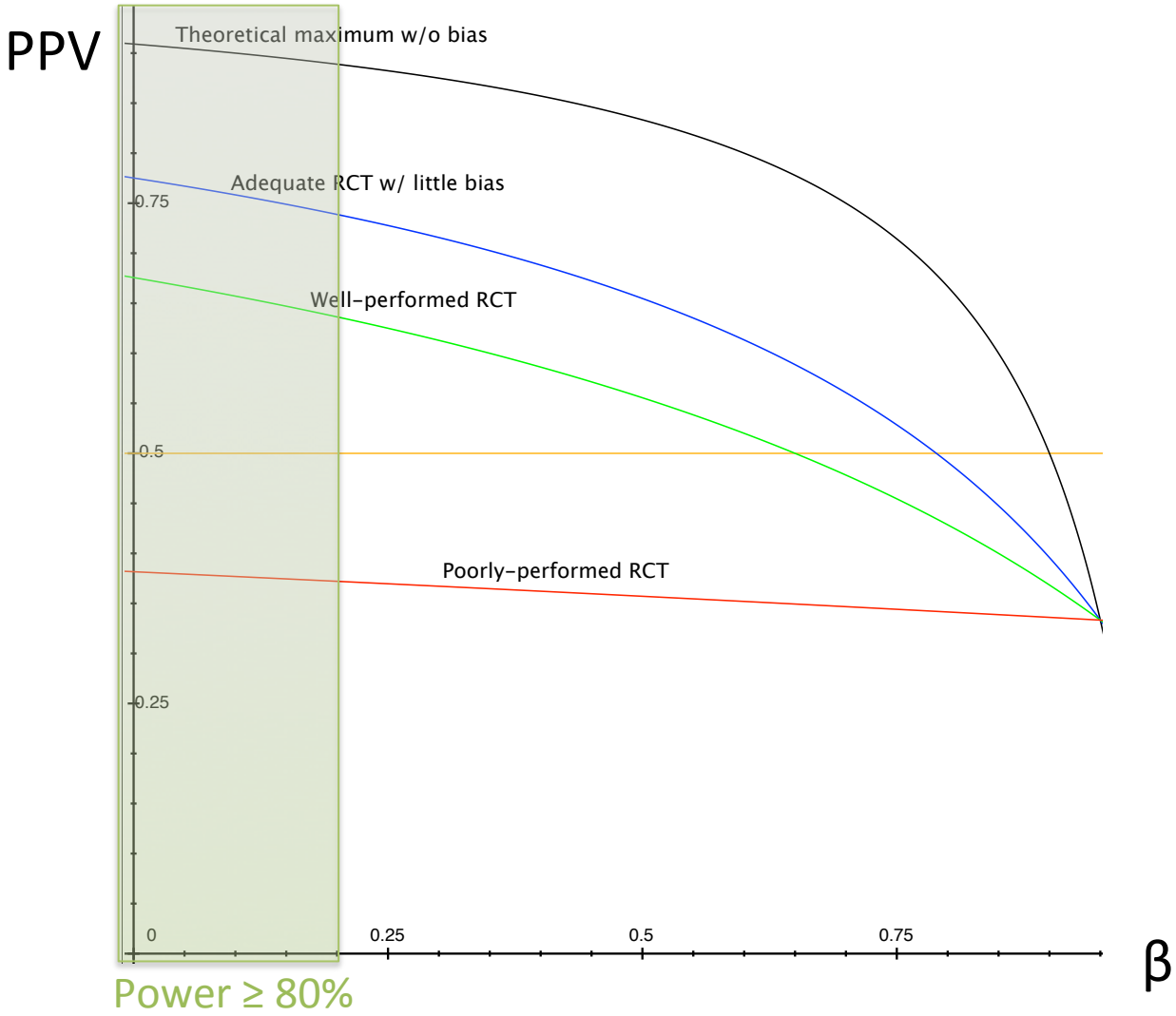
[cf. Ioannidis – Why Most Published Research Findings Are False]

The more biased the study, the less likely the findings are true.



[cf. Ioannidis – Why Most Published Research Findings Are False]

The lower sample size or effect sizes, the less likely the findings are true.



α = .05
R = 1/2

[cf. Ioannidis – Why Most Published Research Findings Are False]

Dependable Empirical Evidence

Dependable Experiment Design
Internal and external validity
Repeatability and Reproducibility

Roy Maxion
Making Experiments Dependable

High Positive Predictive Value

- Adequate power (sample size for effect sizes)
- Few relationships investigated (implies prior)
- Low bias in experiment design and execution
- **Noteworthy result:** PPV ≥ 80%

[Wacholder et al. – Assessing the Probability That a Positive Report is False]

CASE STUDY: COGNITIVE EFFORT OF PASSWORD CHOICE

Is Cognitive Effort Necessary for Strong Passwords?

- **Background:** Cognitive depletion limits human capacity to exert cognitive effort.
- **Aim:** Establish how password choice differs between depleted and non-depleted users.
- **H_1:** *Cognitively depleted users create weaker passwords than non-depleted users.*
- **Method:** Two groups of 50 participants each create a password – one depleted, the other not.

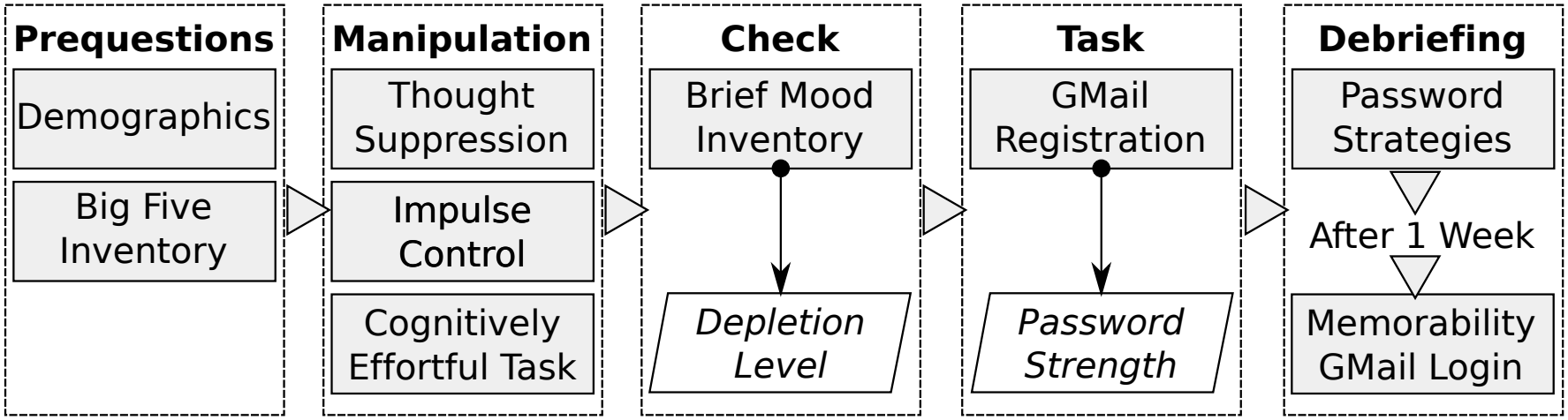
Cognitive Depletion [Baumeister]

- *Cognitive effort* needed to control impulses and to think about 'hard' problems.

- Limited energy to exert cognitive effort.
- Once cognitively depleted, the capacity to exert cognitive effort is incapacitated.

- Beliefs (implicit theories about willpower) impact how much people are depleted. [Job et al. 2010]

Experiment Design



Sampling (N=100)

[A priori power analysis]

Students; international; non-computer science.

Gender: 50 female.

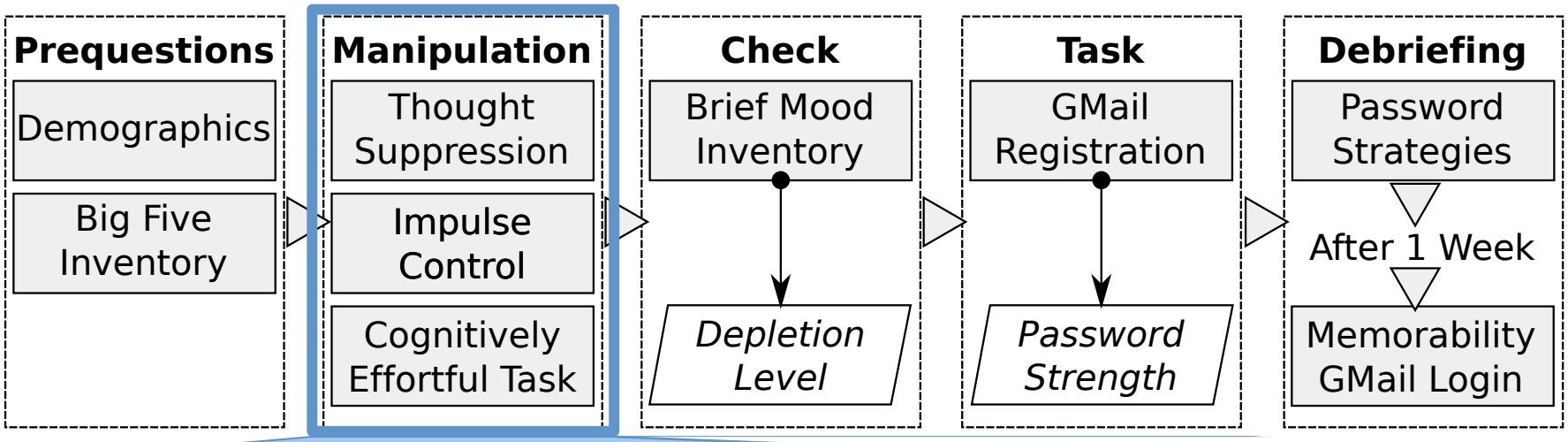
Assignment

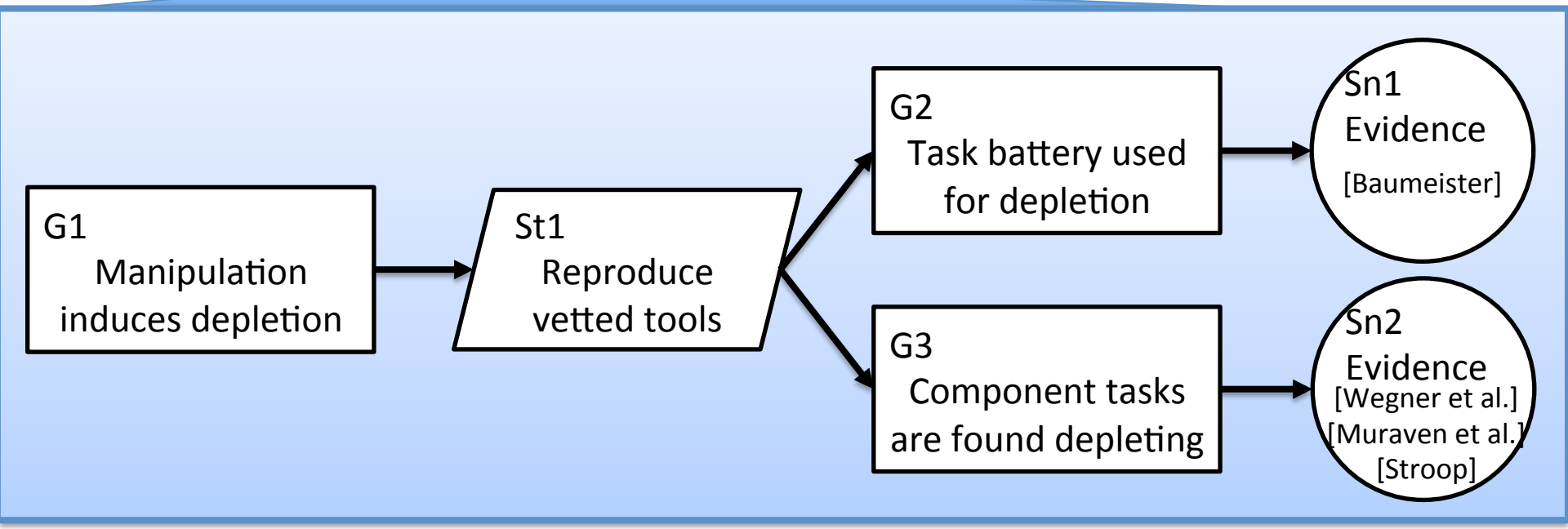
[Constrained random]

Balanced by gender.

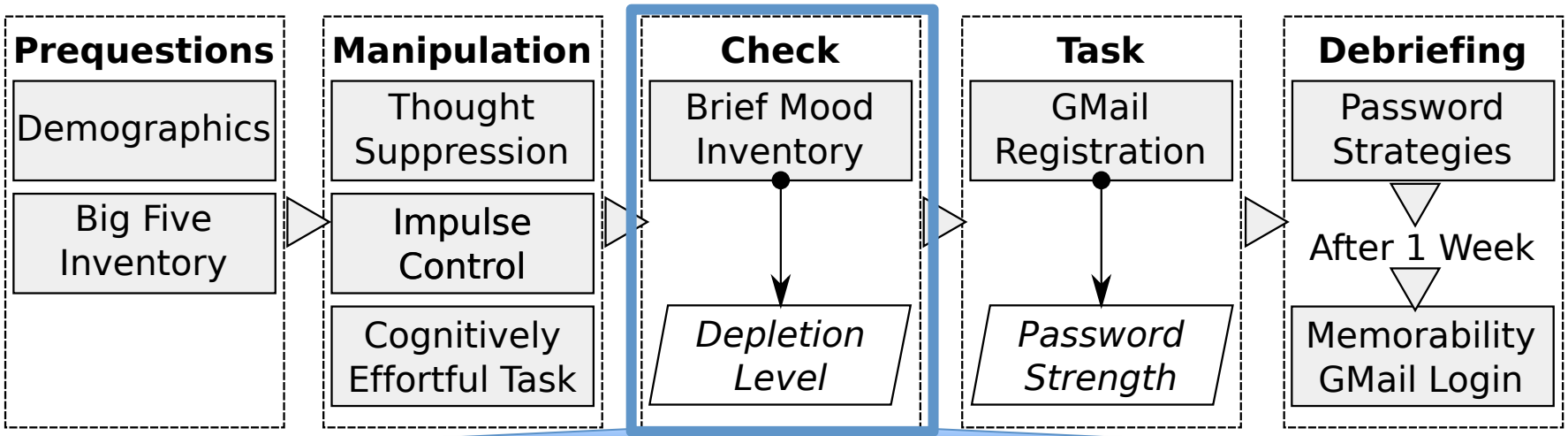
Balanced time-of-day (circadian rhythm).

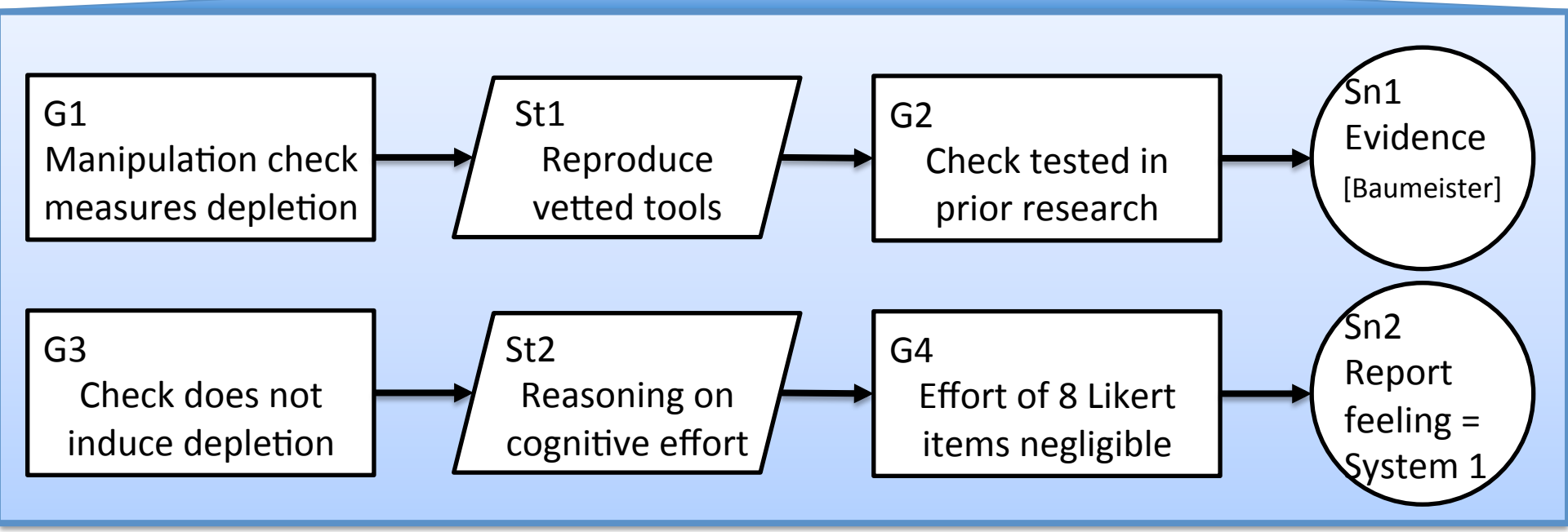
Reasoning on the Manipulation



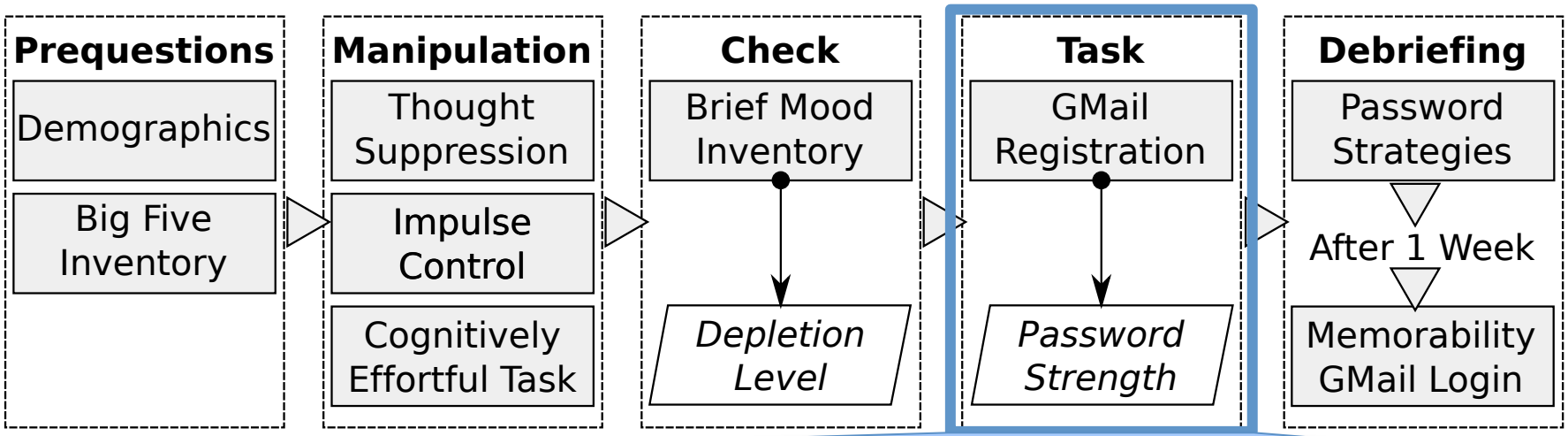


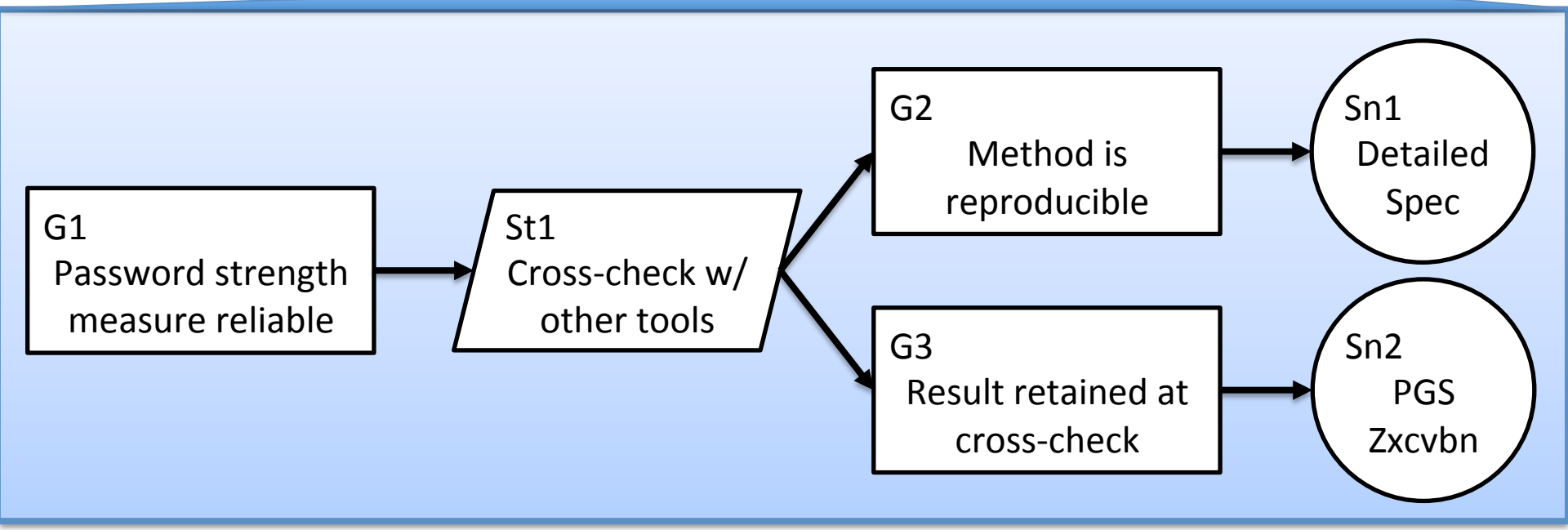
Reasoning on the Manipulation Check





Reasoning on the DV Measurement





Step-Wise Forward Linear Regression

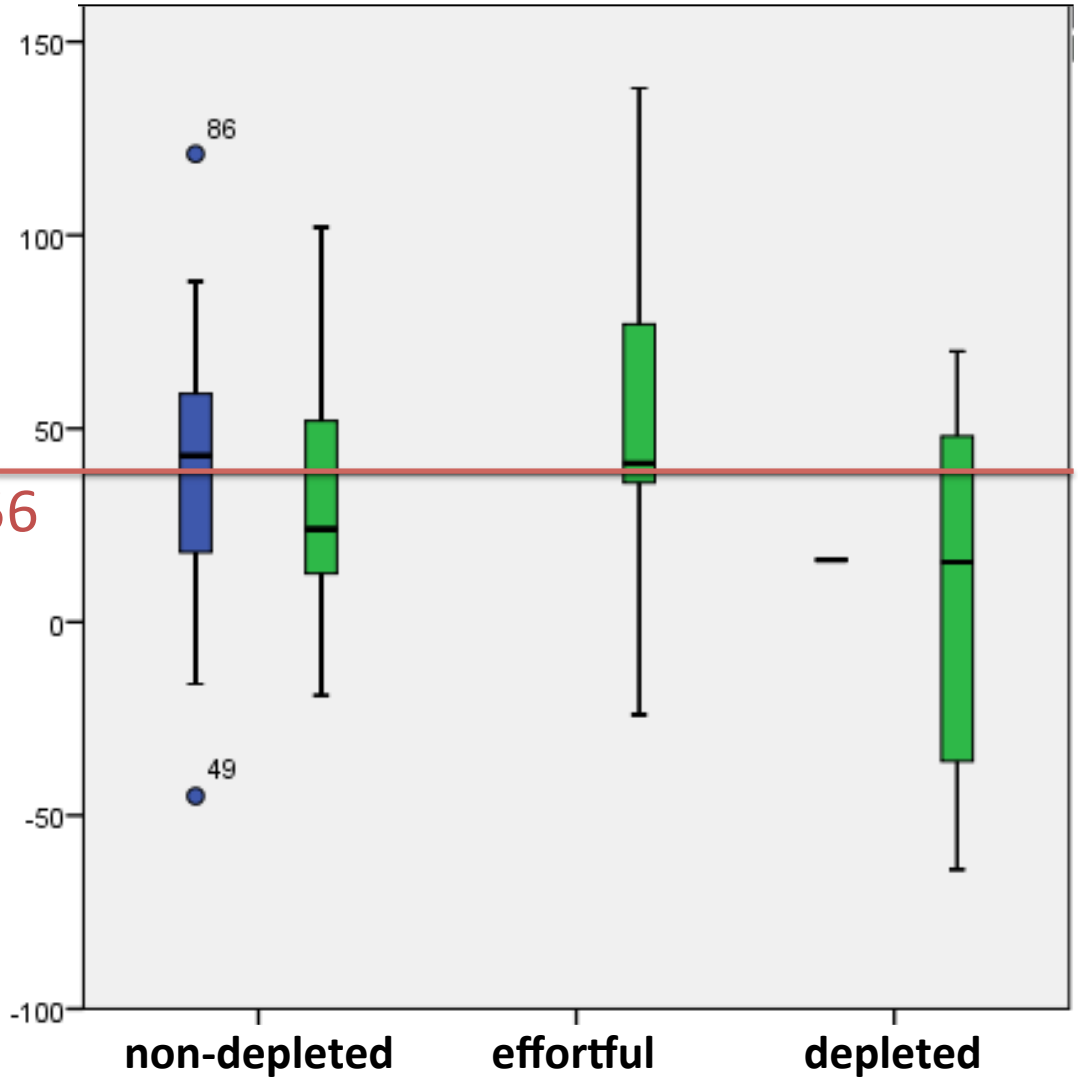
- Adjusted R^2 = 20.6%. Corrected AIC = 711.125.
- Overall effect size: Cohen's f^2 = 0.26.

- **Cognitive depletion** most important predictor (p = .001, importance= 0.371, Cohen's ω^2 = 0.097).
  - **effortful**: coefficient b = 19.03 (p < .001) [.5, 38]
  - **depleted**: coefficient b = -31.62 (p = .006) [-54, -9]
  - **non-depleted:** baseline

- **Moods:** Thoughtfulness and calmness.
- **Big Five:** Agreeableness.

[Not presented in the talk; provided as additional context.]

Password Strength by Depletion and Condition



Control

Experiment

Lessons Learned

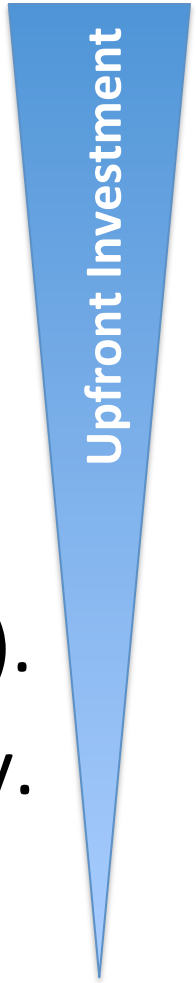
- **Manipulation:** Depletion tasks did not deplete subjects in experiment group equally.
- **Stats:** Unequal sample sizes on depletion levels.
- **Better double-check:** Pretest manipulation, check and measurement components.

For human dimensions of cyber security, it's especially hard to find reliable/vetted manipulation and measurement devices.

OBSERVATIONS: WHAT IT TAKES

Experiments in an Ideal World

- Establish vetted and reproducible manipulation and measurement devices.
- Pretest each component separately and the experiment design overall.
- Determine effect size estimates in pretests.
- Determine sample size (from a priori power).
- Run the experiment reproducibly and validly.
- Follow analysis and reporting standards.



A “Law of Security” for Experiments?

To half your vulnerabilities,

you need to double your expenditure.

(from Adi Shamir’s *Laws of Security*)

To half the biases in an empirical experiment,
you need to double your expenditure.

Unbiased experiments do not exist.

Conclusion

- We need a toolbox and a practice of dependable research methods.
- We need evidence-based reasoning on the validity and low bias of experiments.
- We need assurance of dependable experiments.