# Detecting Answer Copying on Multiple-Choice Tests with Dichotomous Item Scores

Cengiz Zopluoglu
**University of Miami**

Troy T. Chen
Chi-Yu Huang
Andrew Mroch
**ACT, Inc.**

# Answer Copying Indices:

- Person-fit indices vs. Answer similarity indices

- Source of Evidence:
  - Identical incorrect responses
  - Identical correct and incorrect responses
  - All items

- Type of Statistical Distribution
  - Empirical Null distribution
  - Binomial Distribution
  - Poisson Distribution
  - Compound Binomial Distribution
  - Normal Distribution

# Research Purpose

- To investigate the statistical performance of answer copying indices under different simulated conditions by using dichotomous item scores

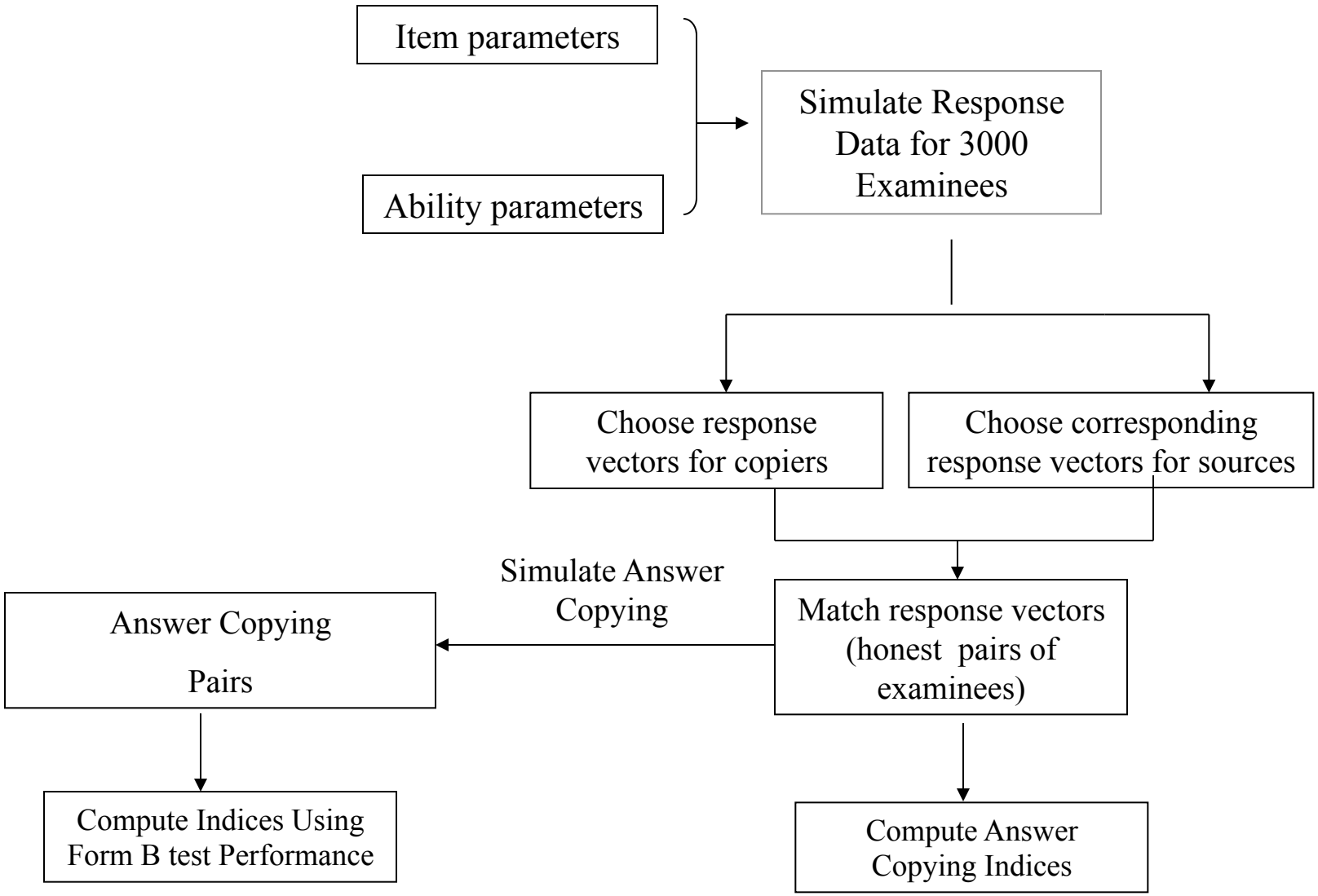
**Independent Variables:**

- **IRT Model:** 2PL and 3PL
- **Test Length:** 30-item and 50-item
- **Ability Group of Pairs**: Low-Low, Low-Medium, Low-High, Medium- Medium, Medium-High, High-High
- **Amount of Copying**: 20%, 40%, 60%
- **Type of Copying:** Random, Random-String

2 x 2 x 6 x 3 x 2 = 144 simulated conditions for power analysis

2 x 2 = 4 conditions for Type I error rate analysis

# Data Generation for One Replication



**Statistical indices included in the study:**

a. Person-fit indices : Lz and Modified Caution Index

b. Answer Similarity Indices: EMRA1, EMRA2, GBT, K and its variants (K1, K2,S1, S2), and $\omega$

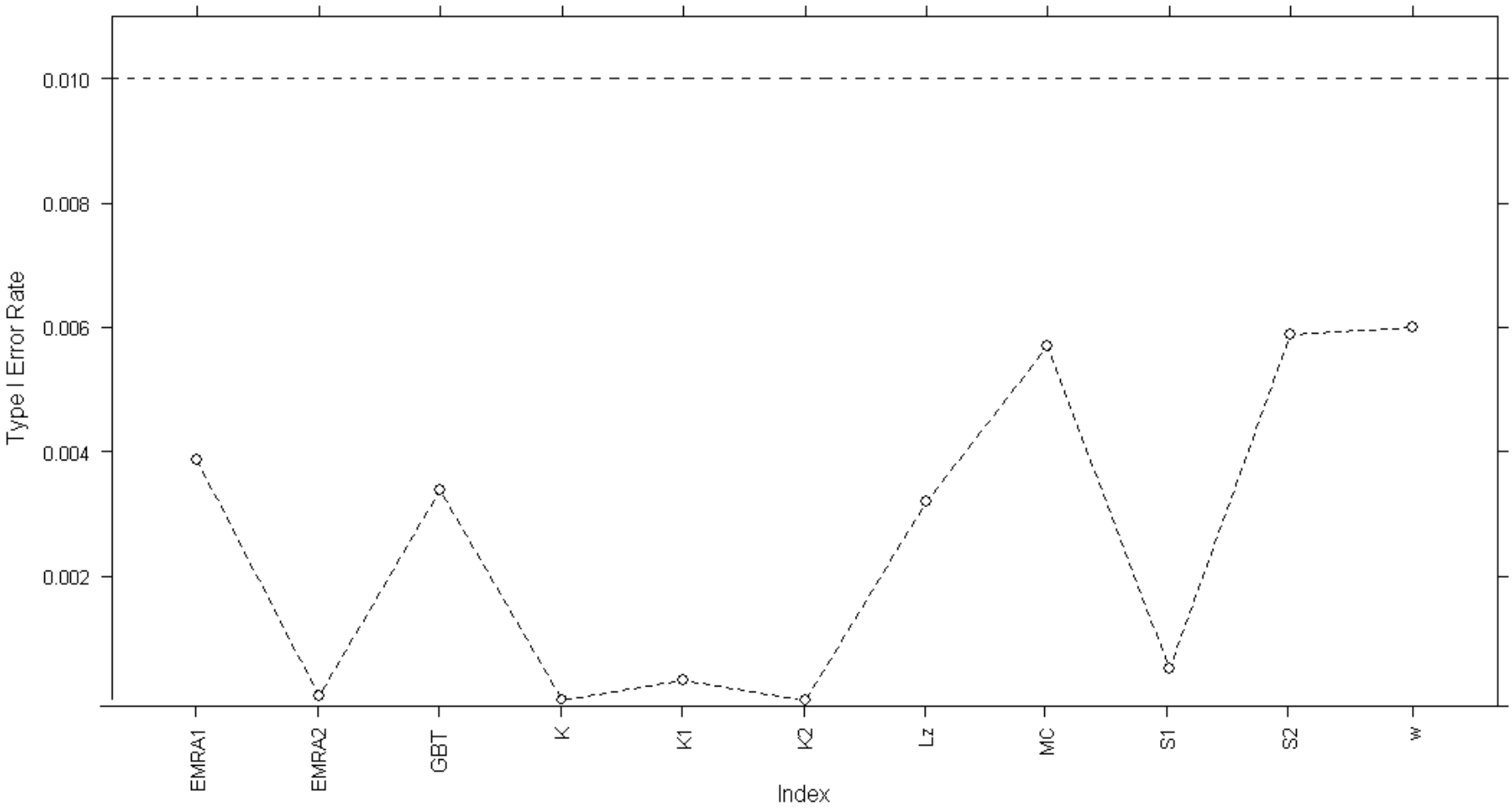
**Analysis:**

**Power:** How many pairs are truly detected out of 5,000 simulated answer copying pairs within each condition by each index at nominal alpha level of .01?

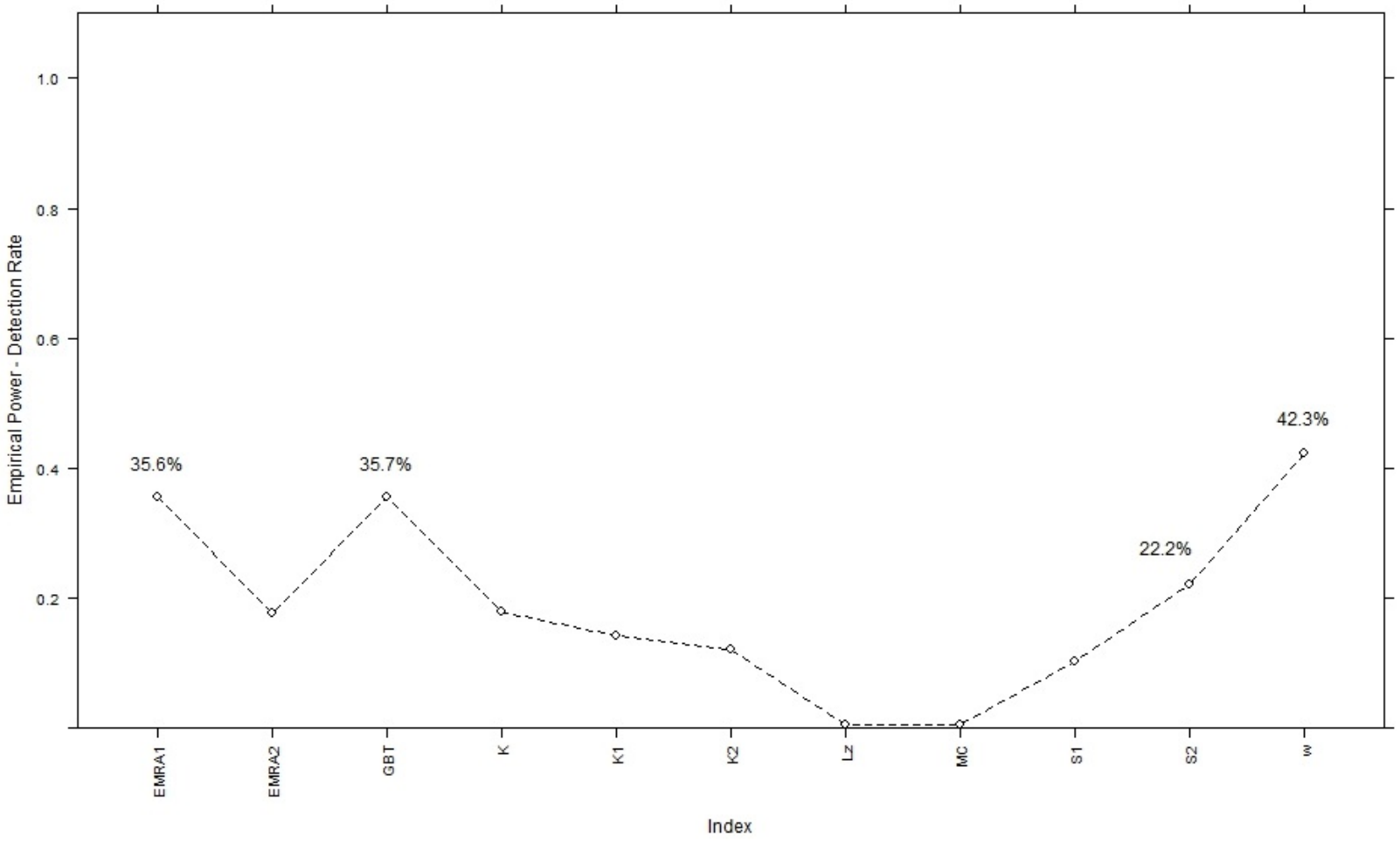
**Type I Error Rate:** How many pairs are falsely detected out of 180,000 simulated honest pairs within each condition by each index at nominal alpha level of .01?

# RESULTS

# Empirical Type I Error Rates at $\alpha = .01$



# Empirical Power at $\alpha = .01$

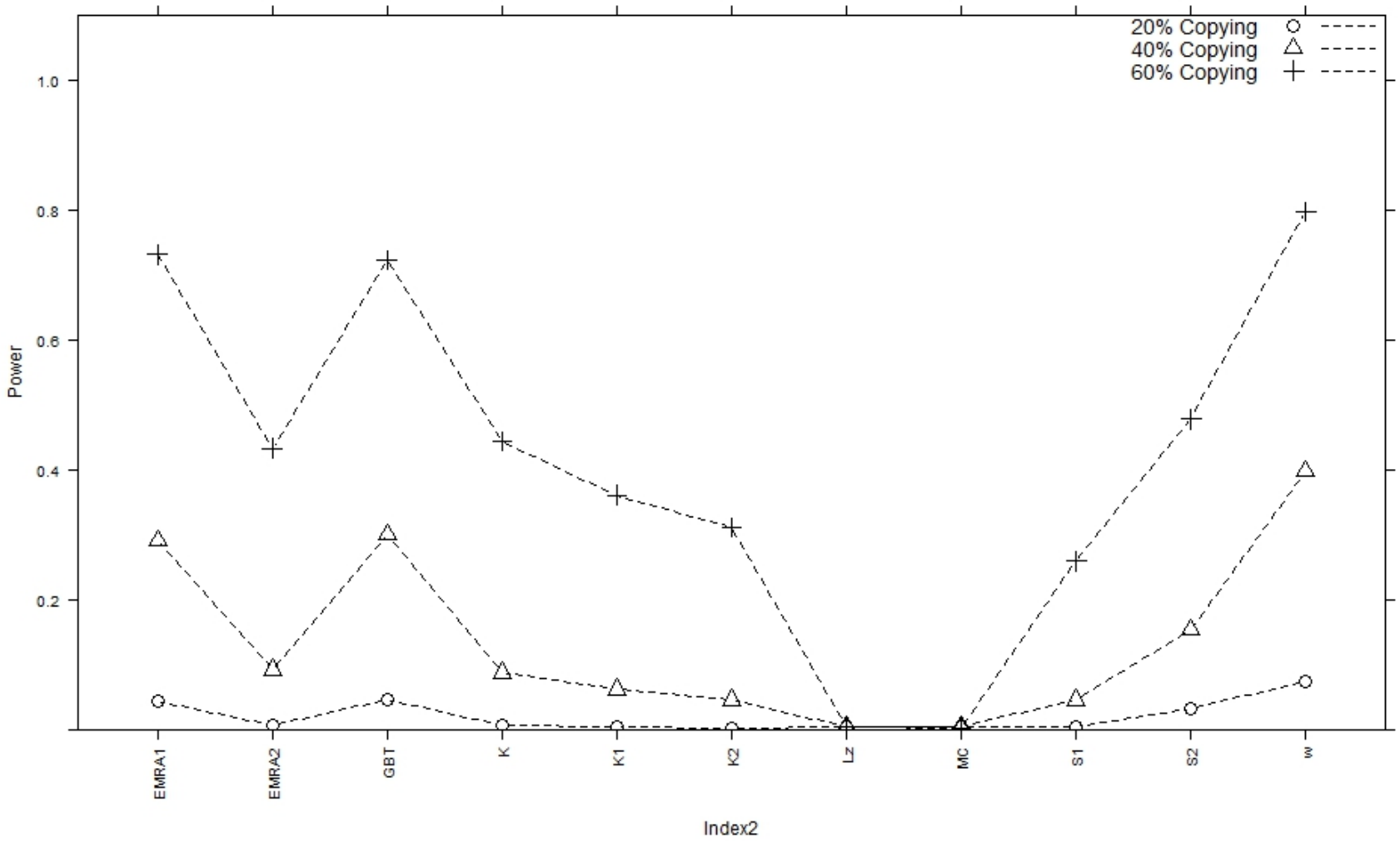


# Eta-Squared Effect Sizes from ANOVA on Statistical Power

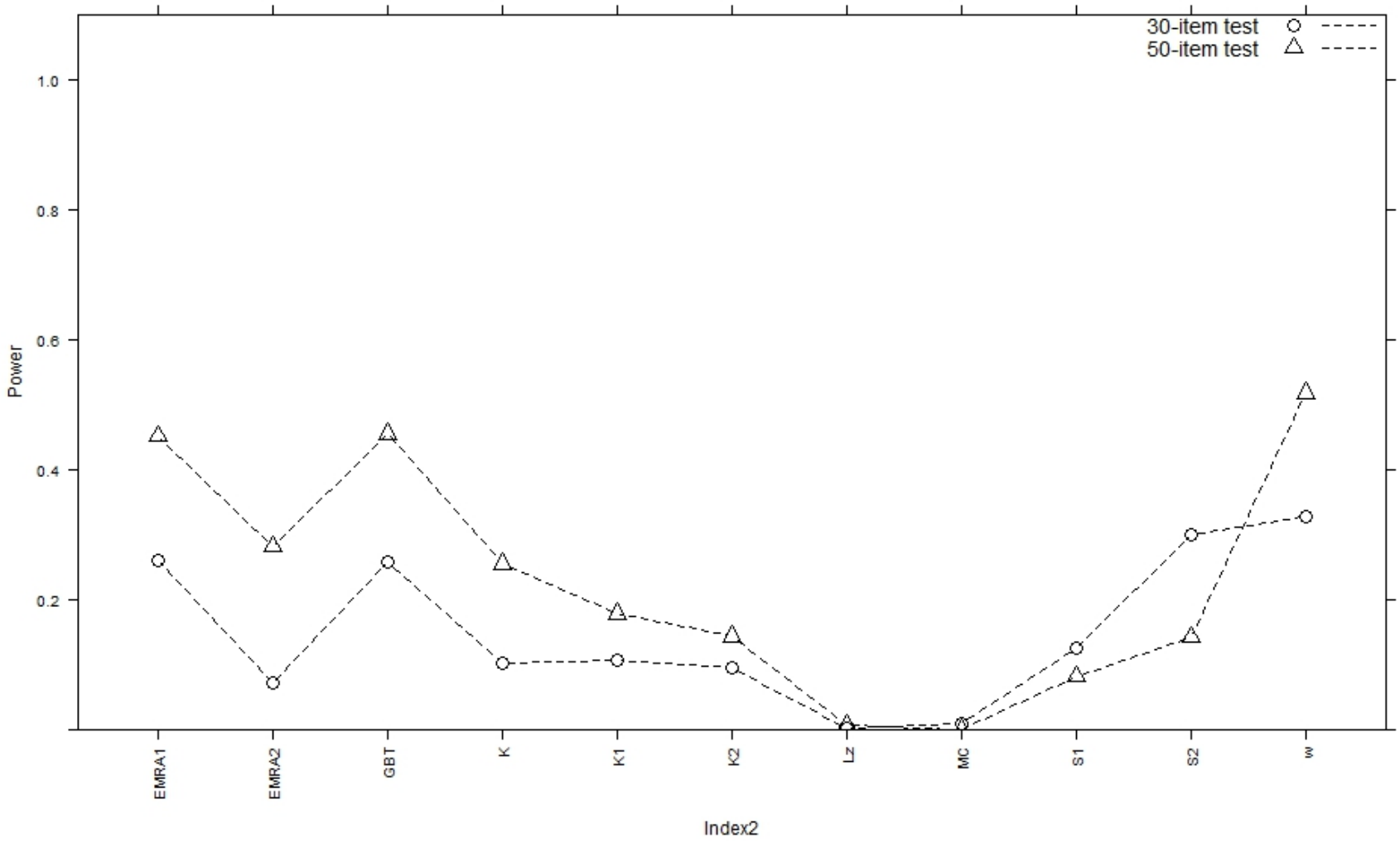
| | EMRA1 | EMRA2 | GBT | K | K1 | K2 | S1 | S2 | w |
|---|---|---|---|---|---|---|---|---|---|
| M | 0.001*** | 0.001*** | 0.001*** | 0.001*** | 0.024*** | 0.024*** | **0.056***** | **0.050***** | 0.003*** |
| L | 0.091*** | **0.181***** | **0.094***** | **0.116***** | 0.046*** | 0.030*** | 0.013*** | **0.082***** | **0.082***** |
| A | **0.818***** | **0.568***** | **0.755***** | **0.726***** | **0.747***** | **0.777***** | **0.610***** | **0.588**** | **0.812***** |
| T | <.001*** | 0.001*** | <.001*** | 0.001*** | 0.002*** | 0.002*** | 0.001*** | <.001** | <.001*** |
| G | 0.032*** | 0.037*** | 0.036*** | 0.014*** | 0.024*** | 0.019*** | **0.096***** | **0.137***** | **0.051***** |

| | |
|---|---|
| M | IRT Model |
| L | Test Length |
| A | Amount of Copying |
| T | Type of Copying |
| G | Ability Group |

# Main effect of Amount of Copying on Empirical Power at $\alpha = .01$



# Main effect of Test Length on Empirical Power at $\alpha = .01$



**Conclusions:**

When dichotomous IRT models and dichotomous response outcomes are used:

1. The $\omega$ index showed highest detection rates, and EMRA1 and GBT also provided reasonable detection rates.

2. The K index and its variants (K1, K2, S1, S2) and EMRA 2 showed relatively lower detection rates

3. Person-fit indices show very low power for detecting answer copying

# What's Next Step?

# Thank you!

**Special Thanks to:**

- Tami Hrasky
- Deborah Harris
- Karen Zimmermann
- Tianli Li