# Replicating Simulation Research: A Case Study

By: Tristan Tibbe

# Overview

- RepliSims Project
- MacKinnon et al. (2004)
- Replication Process
- Factors that Hindered Replication
- Factors that Facilitated Replication
- Recommendations to Improve Replicability

# RepliSims Project

Published in Royal Society Open Science (Luijken et al., 2024)

8 teams of researchers

8 replicated studies

Criteria:

- Published after 2000
- Greater than 1000 citations

# MacKinnon et al. (2004)

Compared confidence intervals (CIs) for the indirect effect constructed using 9 methods:

- z critical values
- M critical values
- Empirical M method
- Jackknife
- Percentile bootstrap
- Bias-corrected bootstrap
- Bootstrap-*t*
- Boostrap-Q
- Monte Carlo

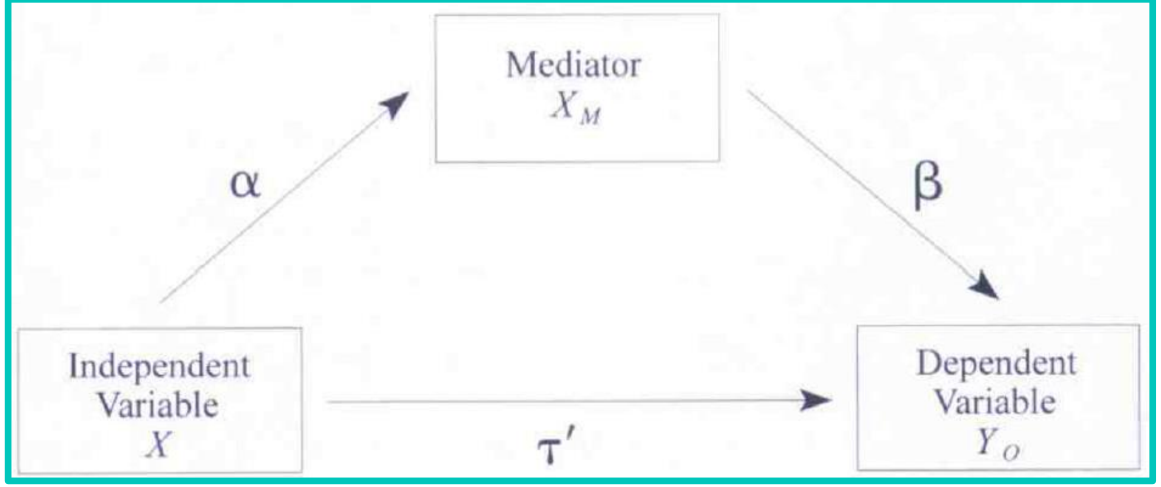


NIH Public Access
**Author Manuscript**
*Multivariate Behav Res*. Author manuscript; available in PMC 2010 February 12.

Published in final edited form as:
*Multivariate Behav Res*. 2004 January 1; 39(1): 99. doi:10.1207/s15327906mbr3901_4.

**Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods**

**David P. MacKinnon**, **Chondra M. Lockwood**, and **Jason Williams**
Arizona State University



# MacKinnon et al. (2004)

Outcomes:

Type I error rate/power

CI balance/width

Findings:

***M* critical values** result in more balanced CIs than **z critical values**

In order of increasing type I error rate/power and decreasing CI width:

**z critical values**
**Jackknife**

→

**Percentile bootstrap**
**Bootstrap-*t***
**Monte Carlo**

→

***M* critical values**
**Empirical *M* method**
**Bootstrap-Q**

→

**Bias-corrected bootstrap**

# Replication Process

1.) Recreate data generating process

2.) Recreate methods

3.) Rerun simulation

4.) Compare results to original findings

# 1.) Recreate Data Generating Process

Manipulated Factors

Sample sizes

Effect sizes

Fixed Factors

with current time as the seed for each simulation. Five different sample sizes corresponding to sample sizes in the social sciences were simulated: 50, 100, 200, 500, and 1000. The

demands of simulation studies of resampling methods. The ten combinations were $\alpha = 0$ $\beta = 0$, $\alpha = 0$ $\beta = .14$, $\alpha = 0$ $\beta = .39$, $\alpha = 0$ $\beta = .59$, $\alpha = \beta = .14$, $\alpha = \beta = .39$, $\alpha = \beta = .59$, $\alpha = .14$ $\beta = .39$. $\alpha = .14$ $\beta = .59$, and $\alpha = .39$ $\beta = .59$. These ten parameter combinations are the ones presented

simulations indicated no difference in power calculations as the direct effect ($\tau'$) increased, so for simplicity the direct effect was always set equal to zero.

the statistical simulations. The data were simulated using Equations 2 and 3, with sample values of $X$, $\varepsilon_2$, and $\varepsilon_3$ generated from a standard normal distribution using the SAS RANNOR function

# 2.) Recreate Methods

$$s_{jackknife} = \sqrt{\frac{N-1}{N}\sum[\theta_{(i)} - \theta_{(\cdot)}]^2} \quad (8)$$

Equations

Programs

Specific Settings

Confidence levels

in increments of .2. These additional values were obtained with a FORTRAN algorithm written by Alan Miller which is a minor modification of the method in Meeker and Escobar (1994)

standard error were used to compute confidence limits as described below. Confidence limits for the indirect effect were calculated for 80%, 90%, and 95% intervals. The proportion of

# Replication Process: Summary of Steps 1-2

| Study 2: Simulation factor | No. levels | Levels |
|---|---|---|
| *Varied* | | |
| Confidence interval method | 9 | $z$ method, $M$ method, empirical-$M$ method, jackknife, percentile bootstrap, bias-corrected bootstrap, bootstrap-$t$, bootstrap-$Q$, Monte Carlo method |
| Sample size | 4 | 25, 50, 100, 200 |
| $\alpha$ effect size | 4 | 0, .14, .39, .59 |
| $\beta$ effect size | 4 | 0, .14, .39, .59 |
| Confidence level | 3 | 95%, 90%, 80% |
| *Fixed* | | |
| Direct effect size | | 0 |
| Intercepts | | 0 |
| *Randomly Sampled* | | |
| $X$ values | | sampled from $N(0, 1)$ |
| Error terms | | sampled from $N(0, 1)$ |

# 3.) Rerun Simulation

Software

Iterations

Seed Values

**Simulation Description**—The SAS® (1989) programming language was used to conduct the statistical simulations. The data were simulated using Equations 2 and 3, with sample values

in the Tables for Study 1. Third, one thousand replications were conducted for each of the 40 combinations of sample size and parameters. Fourth, for each of the 40,000 (4 combinations

of $X$, $\varepsilon_2$, and $\varepsilon_3$ generated from a standard normal distribution using the SAS RANNOR function with current time as the seed for each simulation. Five different sample sizes corresponding to

# 4.) Compare Results to Original Findings

Recreate tables/figures

Find metrics to compare results/identify differences

Bradley's (1978) Liberal Robustness Criterion (0.0125 – 0.0375)

Proportion of True Value to the Left and Right of 95% Confidence Intervals, study 2

| | | Sample Size | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 25 | | 50 | | 100 | | 200 | |
| Indirect Effect | Method | left | right | left | right | left | right | left | right |
| Null Models | *z* | 0.0020* | 0.0028* | 0.0055* | 0.0043* | 0.0090* | 0.0083* | 0.0098* | 0.0078* |
| | *M* | 0.0103* | 0.0140 | 0.0113* | 0.0145 | 0.0180 | 0.0188 | 0.0183 | 0.0130 |
| | Empirical-*M* | 0.0098* | 0.0140 | 0.0128 | 0.0150 | 0.0188 | 0.0195 | 0.0188 | 0.0140 |
| | Jackknife | 0.0033* | 0.0033* | 0.0053* | 0.0063* | 0.0080* | 0.0083* | 0.0103* | 0.0090* |
| | Bootstrap percentile | 0.0090* | 0.0113* | 0.0140 | 0.0150 | 0.0188 | 0.0190 | 0.0195 | 0.0150 |
| | Bootstrap Bias-corrected | 0.0245 | 0.0268 | 0.0255 | 0.0260 | 0.0293 | 0.0330 | 0.0275 | 0.0275 |
| | Bootstrap-*t* | 0.0065* | 0.0088* | 0.0133 | 0.0105* | 0.0160 | 0.0180 | 0.0178 | 0.0138 |
| | Bootstrap-*Q* | 0.0075* | 0.0103* | 0.0125 | 0.0110* | 0.0165 | 0.0183 | 0.0175 | 0.0135 |
| | Monte Carlo | 0.0070* | 0.0108* | 0.0103* | 0.0113* | 0.0165 | 0.0153 | 0.0160 | 0.0110* |
| Non-zero Models | *z* | 0.0030* | 0.0547* | 0.0077* | 0.0577* | 0.0098* | 0.0598* | 0.0132 | 0.0480* |
| | *M* | 0.0120* | 0.0502* | 0.0200 | 0.0467* | 0.0192 | 0.0492* | 0.0198 | 0.0398* |
| | Empirical- *M* | 0.0118* | 0.0408* | 0.0192 | 0.0473* | 0.0190 | 0.0487* | 0.0190 | 0.0378* |
| | Jackknife | 0.0057* | 0.0528* | 0.0072* | 0.0570* | 0.0125 | 0.0582* | 0.0135 | 0.0487* |
| | Bootstrap percentile | 0.0127 | 0.0438* | 0.0187 | 0.0437* | 0.0233 | 0.0413* | 0.0222 | 0.0400* |
| | Bootstrap Bias-corrected | 0.0207 | 0.0553* | 0.0268 | 0.0498* | 0.0288 | 0.0430* | 0.0273 | 0.0340 |
| | Bootstrap-*t* | 0.0098* | 0.0352 | 0.0177 | 0.0372 | 0.0202 | 0.0357 | 0.0223 | 0.0350 |
| | Bootstrap-*Q* | 0.0185 | 0.0603* | 0.0273 | 0.0470* | 0.0297 | 0.0470* | 0.0265 | 0.0365 |
| | Monte Carlo | 0.0098* | 0.0295 | 0.0172 | 0.0317 | 0.0168 | 0.0350 | 0.0182 | 0.0335 |

| | | Sample Size | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 25 | | 50 | | 100 | | 200 | |
| Indirect Effect | Method | left | right | left | right | left | right | left | right |
| Null Models | *z* | 0.0018* | 0.0023* | 0.0033* | 0.0050* | 0.0078* | 0.0083* | 0.0133 | 0.0108* |
| | *M* | 0.013 | 0.0128 | 0.0125 | 0.0158 | 0.016 | 0.0155 | 0.0193 | 0.0158 |
| | Jackknife | 0.0048* | 0.0045* | 0.0043* | 0.0035* | 0.0085* | 0.0085* | 0.0120* | 0.0108* |
| | Bootstrap percentile | 0.0113* | 0.0088* | 0.0125 | 0.0148 | 0.014 | 0.0143 | 0.0178 | 0.0163 |
| | Bootstrap Bias-corrected | 0.0195 | 0.0175 | 0.0243 | 0.0268 | 0.0235 | 0.0255 | 0.0245 | 0.023 |
| | Bootstrap-*t* | 0.019 | 0.0198 | 0.022 | 0.024 | 0.0233 | 0.0255 | 0.0263 | 0.0228 |
| | Bootstrap-*Q* | 0.019 | 0.0198 | 0.022 | 0.024 | 0.0233 | 0.0255 | 0.0263 | 0.0228 |
| | Monte Carlo | 0.0095* | 0.0090* | 0.0095* | 0.0128 | 0.0135 | 0.0145 | 0.019 | 0.0138 |
| Non-zero Models | *z* | 0.0058* | 0.0595* | 0.0077* | 0.0552* | 0.0103* | 0.0553* | 0.0103* | 0.0483* |
| | *M* | 0.0142 | 0.0575* | 0.0152 | 0.0482* | 0.0173 | 0.0487* | 0.0173 | 0.0363 |
| | Jackknife | 0.0093* | 0.0605* | 0.0093* | 0.0550* | 0.0117* | 0.0577* | 0.0108* | 0.0487* |
| | Bootstrap percentile | 0.0142 | 0.0375 | 0.0158 | 0.0397* | 0.0187 | 0.0362 | 0.0173 | 0.0338 |
| | Bootstrap Bias-corrected | 0.0228 | 0.0468* | 0.0235 | 0.0428* | 0.0243 | 0.0378* | 0.0228 | 0.0277 |
| | Bootstrap-*t* | 0.025 | 0.1012* | 0.022 | 0.0818* | 0.0242 | 0.0667* | 0.0233 | 0.0435* |
| | Bootstrap-*Q* | 0.025 | 0.1012* | 0.022 | 0.0818* | 0.0242 | 0.0667* | 0.0233 | 0.0435* |
| | Monte Carlo | 0.0117* | 0.0308 | 0.0128 | 0.0333 | 0.0167 | 0.0313 | 0.0167 | 0.0312 |

Study 2 proportion of true values to left and right of 95 percent confidence intervals

# Replication Process: Summary of Step 4

Findings:

***M* critical values** result in more balanced CIs than **z critical values**

In order of increasing type I error rate/power and decreasing CI width:

**z critical values**
**Jackknife**

→

**Percentile bootstrap**
**Bootstrap-*t***
**Monte Carlo**

→

***M* critical values**
**Empirical *M* method**
**Bootstrap-Q**

→

**Bias-corrected bootstrap**

# Replication Process: Summary of Step 4

Findings:

- ***M* critical values** result in more balanced CIs than **z critical values** ✓
- In order of increasing type I error rate/power and decreasing CI width:

**z critical values**
**Jackknife**

→

**Percentile bootstrap**
**Bootstrap-*t***
**Monte Carlo**

→

***M* critical values**
**Empirical *M* method**
**Bootstrap-Q**

→

**Bias-corrected bootstrap**

# Replication Process: Summary of Step 4

Findings:

- ***M* critical values** result in more balanced CIs than ***z* critical values** ✓

- In order of increasing type I error rate/power and decreasing CI width:

***z* critical values**
**Jackknife**
→
**Percentile bootstrap**
**Bootstrap-*t***
**Monte Carlo**
→
***M* critical values**
**Empirical *M* method**
**Bootstrap-Q**
→
**Bias-corrected bootstrap**

# Replication Process: Summary of Step 4

Findings:

- ***M* critical values** result in more balanced CIs than **z critical values** ✓
- In order of increasing type I error rate/power and decreasing CI width:

**z critical values**
**Jackknife**

→

**Percentile bootstrap**
**Bootstrap-*t***
**Monte Carlo**

→

***M* critical values**
**Bootstrap-Q**

→

**Bias-corrected bootstrap**

# Replication Process: Summary of Step 4

Findings:

- ***M* critical values** result in more balanced CIs than **z critical values** ✓
- In order of increasing type I error rate/power and decreasing CI width:

**z critical values**
**Jackknife**

→

**Percentile bootstrap**
**Bootstrap-*t***
**Monte Carlo**

→

***M* critical values**
**Bootstrap-Q**

→

**Bias-corrected bootstrap**

# Replication Process: Summary of Step 4

Findings:

- ***M* critical values** result in more balanced CIs than **z critical values**
- In order of increasing type I error rate/power and decreasing CI width:

**z critical values**
**Jackknife**

→

**Percentile bootstrap**
**Bootstrap-*t*/-*Q***
**Monte Carlo**

→

***M* critical values**

→

**Bias-corrected bootstrap**

# Factors that Hindered Replication

Broken/Outdated Links

Unclear Information

- Methods Implementation
- Error Handling

Paywalls

[2]The empirical-*M* critical values are given at our website given in Footnote 1.

Equations 6 and 7 were used to calculate the *M* confidence limits. The upper and lower critical values were obtained from the table in Meeker et al. (1981) for percentiles of .025 and .975.

**Bootstrap-Q:** The bootstrap-*Q* is a transformation of the bootstrap-*t* that makes the distribution more closely follow the *t* distribution (Manly, 1997). The bootstrap-*Q* is obtained by transforming the bootstrap-*t* using Equation 9 shown below where *s* is skewness in each bootstrap distribution of *T*, *T* is the bootstrap-*t* value in each individual bootstrap sample, and *N* is the sample size (Manly, 1997).

$$Q(T)=T+(sT^2)/3+(s^2T^3)/27+s/(6N) \quad (9)$$

Led to increased “replicator degrees of freedom”

# Factors that Facilitated Replication

Explicitly stated simulation conditions

Provided equations or instructions for many methods used

**Simulation Description—**The simulation procedure in Study 1 was used in Study 2 with four exceptions: sample size, parameter combinations, number of replications, and resampling methods. First, only four sample sizes were simulated: 25, 50, 100, and 200. Because resampling methods are particularly useful when sample sizes are small, the two largest sample sizes from Study 1 were dropped and a sample size of 25 was added. Second, a subset of the combinations of parameter values were simulated to reduce the considerable computational demands of simulation studies of resampling methods. The ten combinations were $\alpha = 0$ $\beta = 0$, $\alpha = 0$ $\beta = .14$, $\alpha = 0$ $\beta = .39$, $\alpha = 0$ $\beta = .59$, $\alpha = \beta = .14$, $\alpha = \beta = .39$, $\alpha = \beta = .59$, $\alpha = .14$ $\beta = .39$. $\alpha = .14$ $\beta = .59$, and $\alpha = .39$ $\beta = .59$. These ten parameter combinations are the ones presented in the Tables for Study 1. Third, one thousand replications were conducted for each of the 40 combinations of sample size and parameters. Fourth, for each of the 40,000 (4 combinations of sample size times 10 parameter value combinations times 1000 replications) different data sets, six resampling methods were applied. For the bootstrap methods, a total of 1000 resampled data sets from each of the 40,000 data sets were used. That is, each bootstrap method entailed 1,000,000 (1000 replications times 1000 bootstrap samples) data sets for each of the 40 combinations of sample size and parameter values. For the jackknife method, the number of samples was the same as the sample size (*N*). Each of the resampling methods are described in more detail in the next section.

# Recommendations to Improve Replicability

Be Specific/Explicit About:

- Data generation
- Method implementation
- Error handling
- Results
- Use supplemental material if necessary

Use Permanent Links:

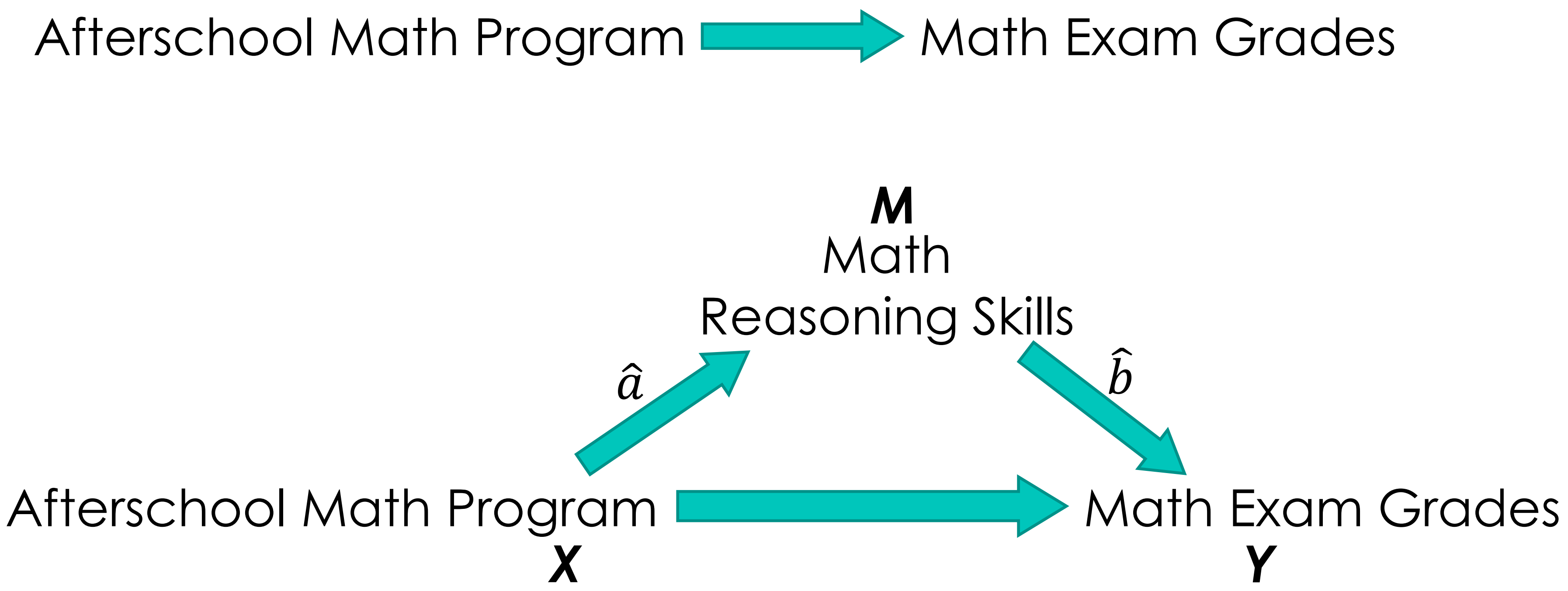
- Upload materials/code to repository (e.g., OSF)

# Questions?

# References

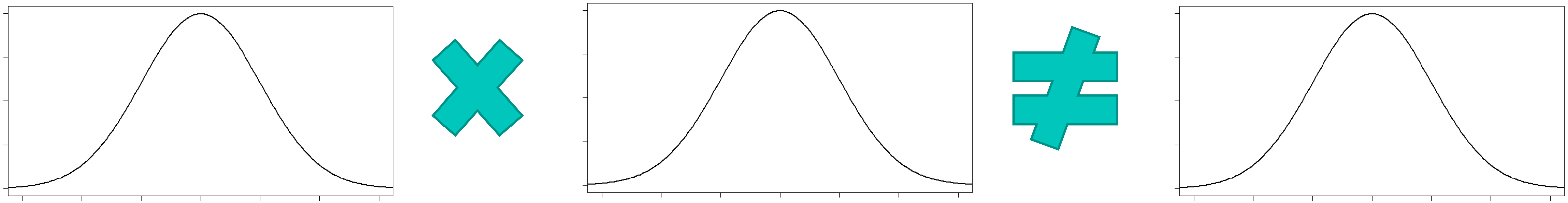

- Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152.
- Luijken, K., Lohmann, A., Alter, U., Claramunt Gonzalez, J., Clouth, F. J., Fossum, J. L., ... & Groenwold, R. H. H. (2024). Replicability of simulation studies for the investigation of statistical methods: The RepliSims project. *Royal Society Open Science, 11*(1).
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*(1), 99-128.
- Manly, B. F. (1997). Randomization, bootstrap and Monte Carlo methods in biology. Chapman and Hall/CRC.
- Meeker, W. Q., Cornwell, L. W., & Aroian, L. A. (1981). Selected tables in mathematical statistics, Vol. VII: The product of two normally distributed random variables. *American Mathematical Society*.
- Meeker, W. Q., & Escobar, L. A. (1994). An algorithm to compute the CDF of the product of two normal random variables. *Communications in Statistics-Simulation and Computation, 23*(1), 271-280.
- SAS Institute. SAS (Version 6.12) [Computer program]. Cary, NC: Author; 1989.

# Mediation Analysis



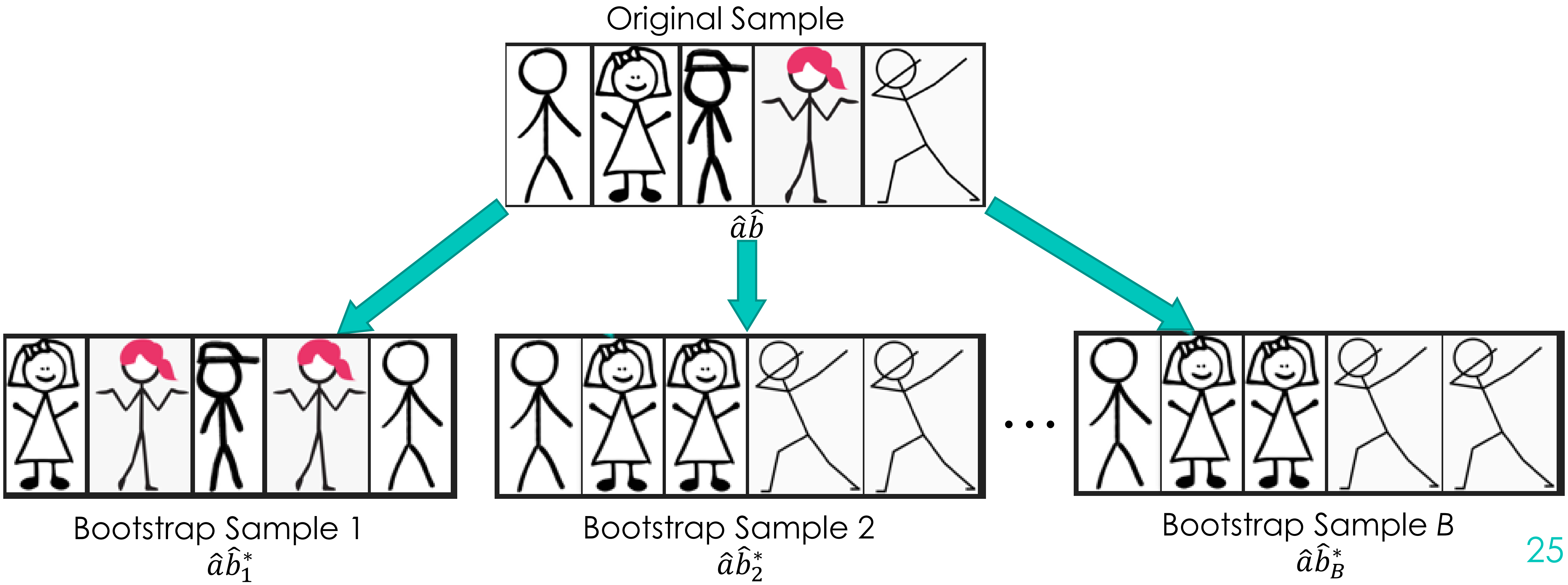
# Indirect Effect in Mediation Analysis

**Distribution of $\hat{a}$** **Distribution of $\hat{b}$** **Distribution of $\hat{a}\,\hat{b}$**



How to test indirect effect for statistical significance?

# Bootstrapping



# Bootstrapping

$\hat{a}\hat{b}_1^*$, $\hat{a}\hat{b}_2^*$, $\dots$, $\hat{a}\hat{b}_B^*$

# Bootstrapping

$\hat{a}\hat{b}_1^*,$ $\hat{a}\hat{b}_2^*,$ $\ldots,$ $\hat{a}\hat{b}_{100}^*$

# Bootstrapping

$\hat{a}\hat{b}_1^*, \quad \hat{a}\hat{b}_2^*, \quad \hat{a}\hat{b}_3^*, \quad \hat{a}\hat{b}_4^*, \quad \hat{a}\hat{b}_5^*, \quad \ldots, \quad \hat{a}\hat{b}_{96}^*, \quad \hat{a}\hat{b}_{97}^*, \quad \hat{a}\hat{b}_{98}^*, \quad \hat{a}\hat{b}_{99}^*, \quad \hat{a}\hat{b}_{100}^*$

# Bootstrapping

$\hat{a}\hat{b}_1^*,\ \hat{a}\hat{b}_2^*,\ \hat{a}\hat{b}_3^*,\ \hat{a}\hat{b}_4^*,\ \hat{a}\hat{b}_5^*,\ \ldots,\ \hat{a}\hat{b}_{96}^*,\ \hat{a}\hat{b}_{97}^*,\ \hat{a}\hat{b}_{98}^*,\ \hat{a}\hat{b}_{99}^*,\ \hat{a}\hat{b}_{100}^*$

100 Bootstrap Indirect Effects

$\hat{a}\hat{b}_{97}^*,\ \hat{a}\hat{b}_3^*,\ \hat{a}\hat{b}_1^*,\ \hat{a}\hat{b}_{100}^*,\ \hat{a}\hat{b}_5^*,\ \ldots,\ \hat{a}\hat{b}_4^*,\ \hat{a}\hat{b}_2^*,\ \hat{a}\hat{b}_{98}^*,\ \hat{a}\hat{b}_{99}^*,\ \hat{a}\hat{b}_{96}^*$

Order from Smallest to Largest

$\hat{a}\hat{b}_{97}^*,\ \hat{a}\hat{b}_3^*\ \Big|\ \hat{a}\hat{b}_1^*,\ \hat{a}\hat{b}_{100}^*,\ \hat{a}\hat{b}_5^*,\ \ldots,\ \hat{a}\hat{b}_4^*,\ \hat{a}\hat{b}_2^*,\ \hat{a}\hat{b}_{98}^*\ \Big|\ \hat{a}\hat{b}_{99}^*,\ \hat{a}\hat{b}_{96}^*$

2.5th percentile

97.5th percentile

95% Percentile Bootstrap Confidence Interval (PBCI)



(Efron & Tibshirani, 1993)

# Bootstrapping

$$\hat{a}\hat{b}^*_{97}, \hat{a}\hat{b}^*_{3} \;\Big|\; \hat{a}\hat{b}^*_{1}, \hat{a}\hat{b}^*_{100}, \hat{a}\hat{b}^*_{5}, \ldots, \hat{a}\hat{b}^*_{4}, \hat{a}\hat{b}^*_{2}, \hat{a}\hat{b}^*_{98} \;\Big|\; \hat{a}\hat{b}^*_{99}, \hat{a}\hat{b}^*_{96}$$

Percentile Bootstrap Confidence Interval (PBCI)

$$\hat{a}\hat{b}^*_{97} \;\Big|\; \hat{a}\hat{b}^*_{3}, \hat{a}\hat{b}^*_{1}, \hat{a}\hat{b}^*_{100}, \hat{a}\hat{b}^*_{5}, \ldots, \hat{a}\hat{b}^*_{4} \;\Big|\; \hat{a}\hat{b}^*_{2}, \hat{a}\hat{b}^*_{98}, \hat{a}\hat{b}^*_{99}, \hat{a}\hat{b}^*_{96}$$

Significance-Tested Bias-Corrected Bootstrap Confidence Interval (stBCBI)

Bias-Corrected Bootstrap Confidence Interval (BCBI)

30% Winsorized Bias-Corrected Bootstrap Confidence Interval (WBCBI)

Reduced Bias-Corrected Bootstrap Confidence Interval (rBCBI)

(Efron & Tibshirani, 1993)
(Chen & Fritz, 2021)
(Stine, 1989)
(Tibbe & Montoya, 2022)