Testing replication of SEM

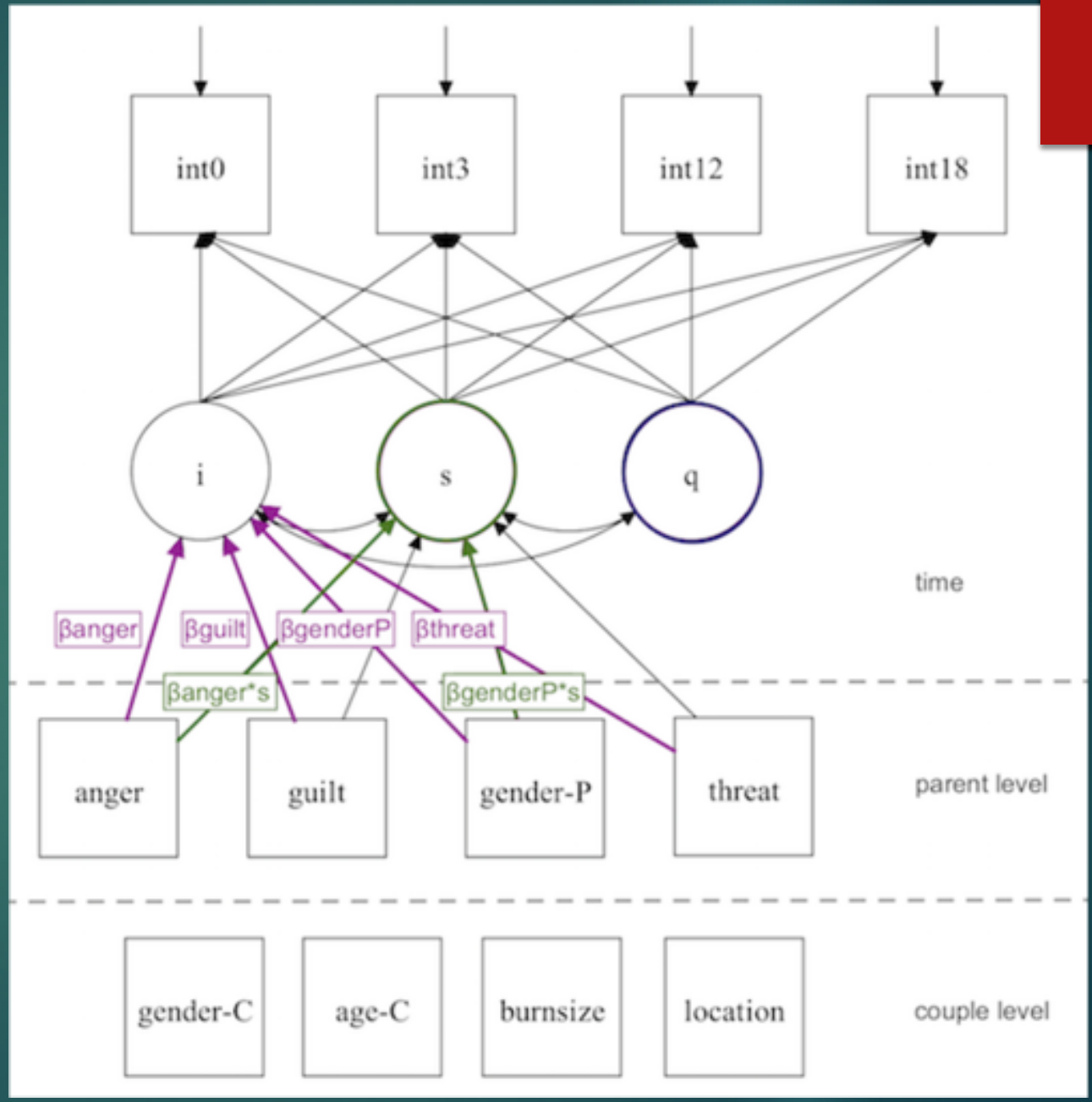
MARIËLLE ZONDERVAN-ZWIJNENBURG

PROMOTOR HERBERT HOIJTINK

Zondervan-Zwijnenburg, M.A.J. (2019). How to Test Replication for Structural Equation Models. *PsyArXiv*. doi: 10.31234/osf.io/uvh5s

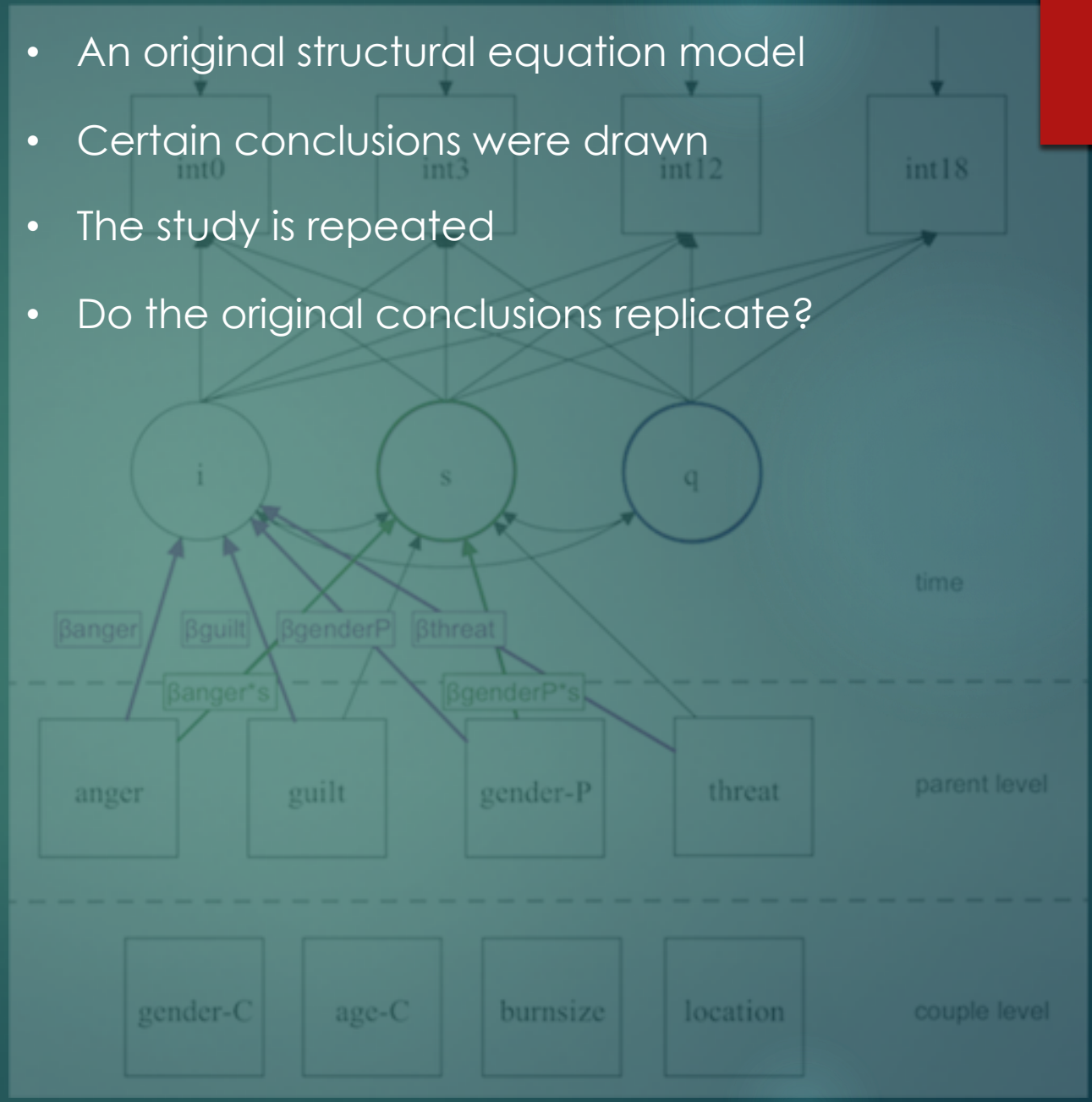
What?

Bakker, A., Van der Heijden, P. G., Van Son, M. J., & Van Loey, N. E. (2013). Course of traumatic stress reactions in couples after a burn event to their young child. *Health Psychology*, *32*(10), 1076–1083. doi: 10.1037/a0033983



What?

- An original structural equation model
- Certain conclusions were drawn
- The study is repeated
- Do the original conclusions replicate?



Bakker, A., Van der Heijden, P. G., Van Son, M. J., & Van Loey, N. E. (2013). Course of traumatic stress reactions in couples after a burn event to their young child. *Health Psychology*, 32(10), 1076–1083. doi: 10.1037/a0033983

Why?

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.188), representing a substantial decline. Ninety-sev… g-inal studies had significant …5). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

ON OUR WEB SITE

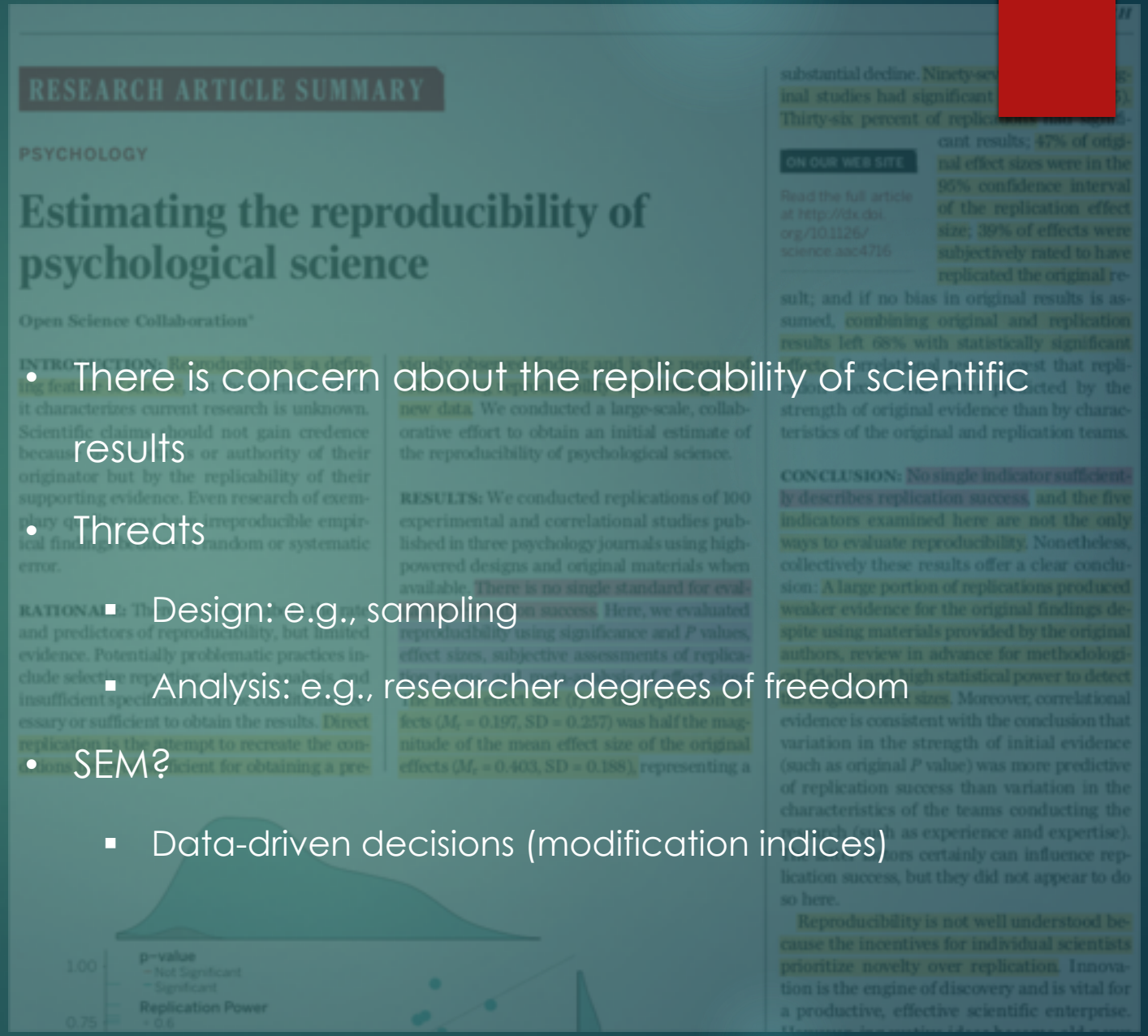
Read the full article at http://dx.doi.org/10.1126/science.aac4716

CONCLUSION: No single indicator sufficiently describes replication success, and the five indicators examined here are not the only ways to evaluate reproducibility. Nonetheless, collectively these results offer a clear conclusion: A large portion of replications produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original *P* value) was more predictive of replication success than variation in the characteristics of the teams conducting the research (such as experience and expertise). The latter factors certainly can influence replication success, but they did not appear to do so here.

Reproducibility is not well understood because the incentives for individual scientists prioritize novelty over replication. Innovation is the engine of discovery and is vital for a productive, effective scientific enterprise.

Why?

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716


RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

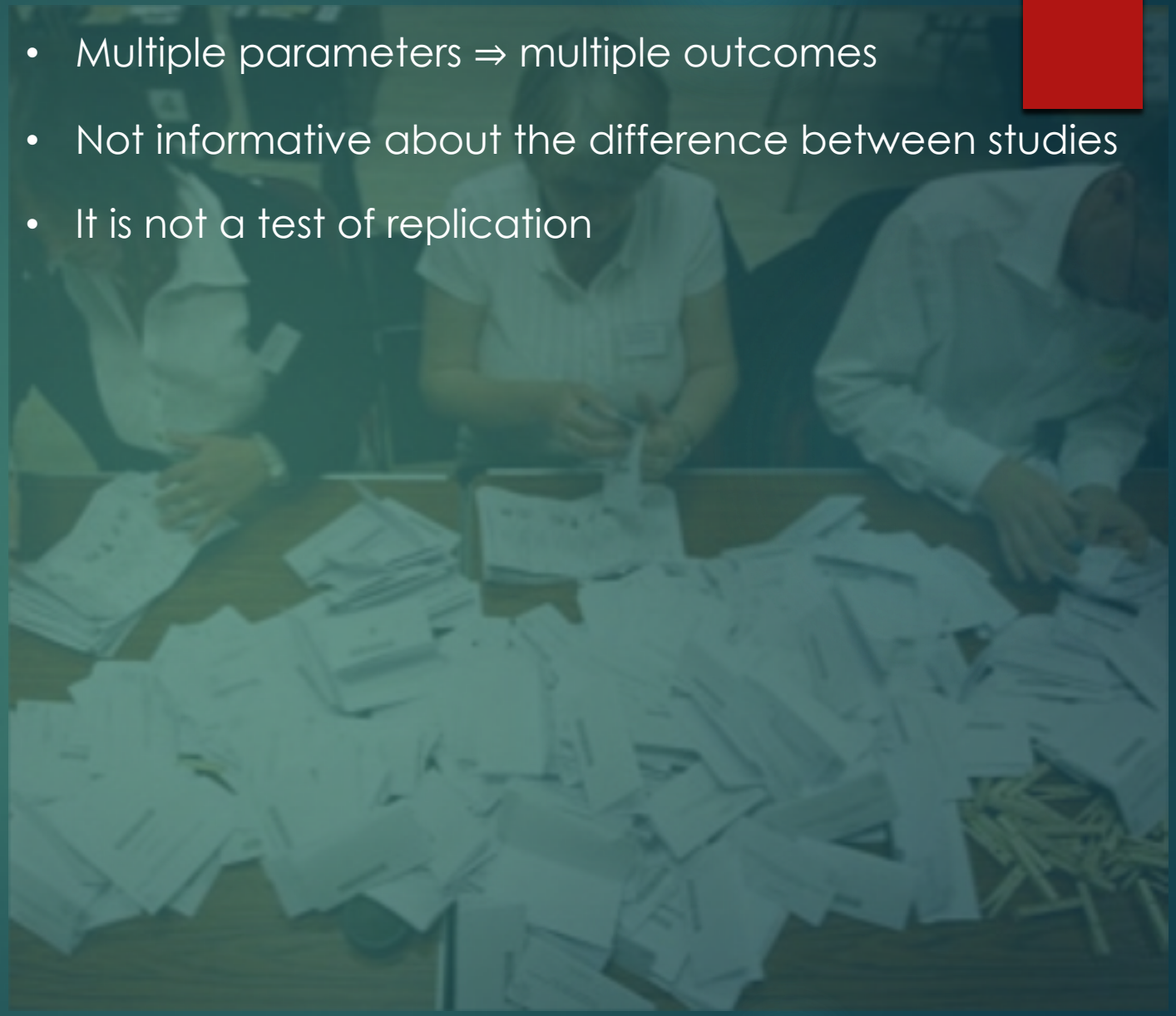
- There is concern about the replicability of scientific results
- Threats
  - Design: e.g., sampling
  - Analysis: e.g., researcher degrees of freedom
- SEM?
  - Data-driven decisions (modification indices)

How?

VOTE COUNTING?

How?

VOTE COUNTING?



- Multiple parameters ⇒ multiple outcomes
- Not informative about the difference between studies
- It is not a test of replication

How?

MODEL FIT?

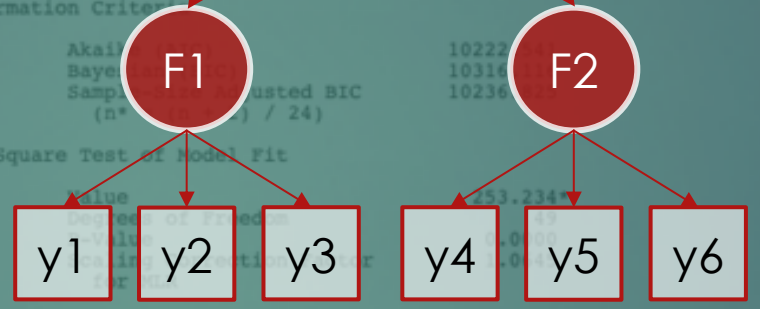
```
THE MODEL ESTIMATION TERMINATED NORMALLY



MODEL FIT INFORMATION

Number of Free Parameters                       25

Loglikelihood

          H0 Value                          -5086.271
          H0 Scaling Correction Factor          1.2713
            for MLR
          H1 Value                          -4951.489
          H1 Scaling Correction Factor          1.1344
            for MLR

Information Criteria

          Akaike (AIC)                      10222.541
          Bayesian (BIC)                    10316.116
          Sample-Size Adjusted BIC          10236.825
            (n* = (n + 2) / 24)

Chi-Square Test of Model Fit

          Value                               253.234*
          Degrees of Freedom                       49
          P-Value                                0.0000
          Scaling Correction Factor              1.0645
            for MLR

*   The chi-square value for MLM, MLMV, MLR, ULSMV, WLSM and WLSMV cannot be used
    for chi-square difference testing in the regular way.  MLM, MLR and WLSM
    chi-square difference testing is described on the Mplus website.  MLMV, WLSMV,
    and ULSMV difference testing is done using the DIFFTEST option.

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                               0.116

CFI/TLI

          CFI                                    0.730
          TLI                                    0.702

Chi-Square Test of Model Fit for the Baseline Model

          Value                                810.404
          Degrees of Freedom                       54
          P-Value                                0.0000

SRMR (Standardized Root Mean Square Residual)

          Value for Within                       0.134
          Value for Between                      0.154
```

How?

MODEL FIT?

- Could be interesting if model fit is the main result of the original study

  (measurement ⇒ factor structure)



- Not a test of replication

How?

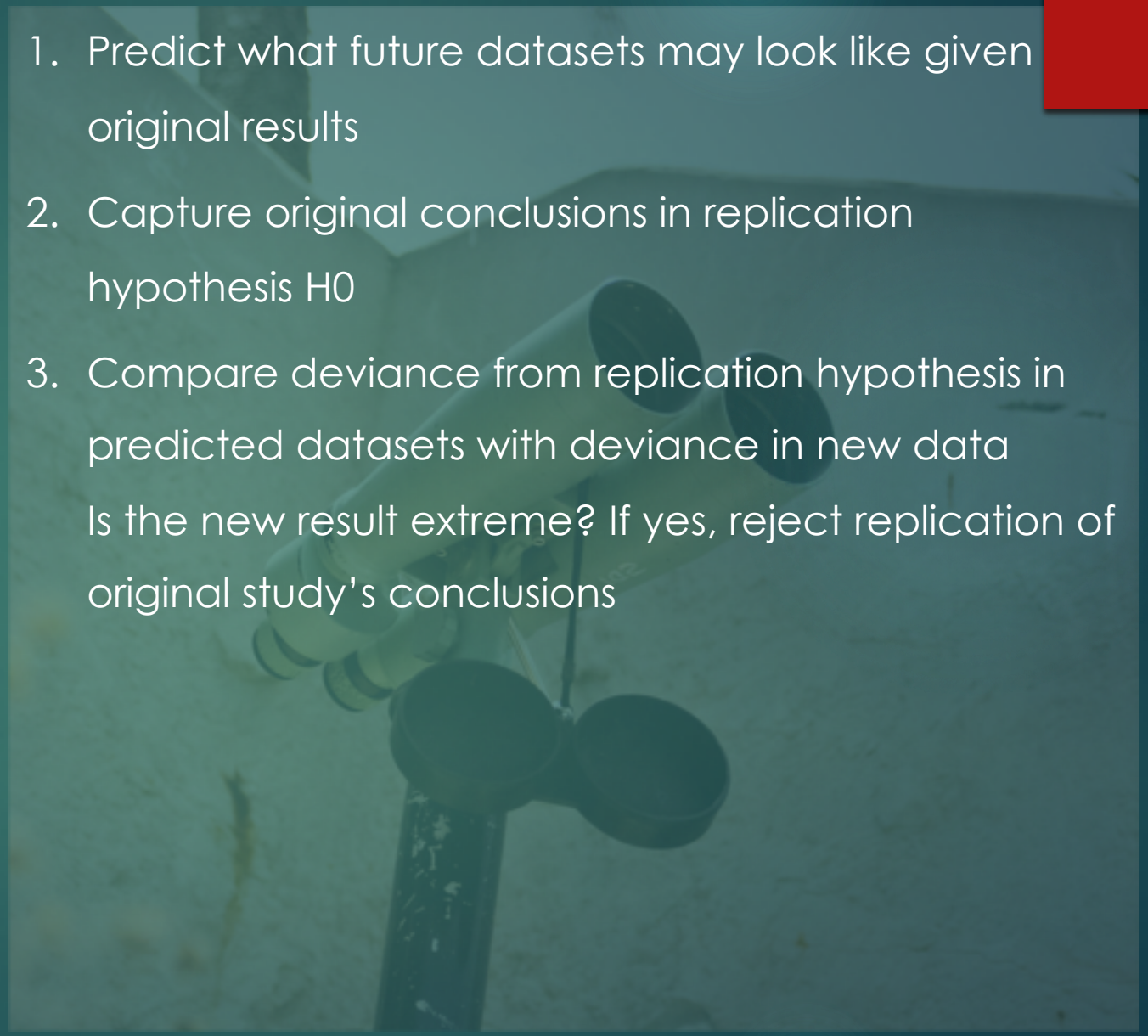
PRIOR PREDICTIVE P-VALUE

Zondervan-Zwijnenburg, M. (2019). How to Test Replication for Structural Equation Models. *PsyArXiv*. doi: 10.31234/osf.io/uvh5s

How?

PRIOR PREDICTIVE P-VALUE

Zondervan-Zwijnenburg, M. (2019). How to Test Replication for Structural Equation Models. *PsyArXiv*. doi: 10.31234/osf.io/uvh5s



1. Predict what future datasets may look like given original results
2. Capture original conclusions in replication hypothesis H0
3. Compare deviance from replication hypothesis in predicted datasets with deviance in new data

   Is the new result extreme? If yes, reject replication of original study's conclusions

1. Predict future datasets

- Predictive distribution

```
install.packages("Replication")          #CRAN R-package
library(Replication)

y.o <- read.table("yo.txt",header=TRUE)    #orig. data
model <- readLines("model.lav")            #model
n.r <- dim(y.r)[2]                         #N new data

step1 <- ppc.step1(y.o=y.o,model=model,n.r=n.r)
```

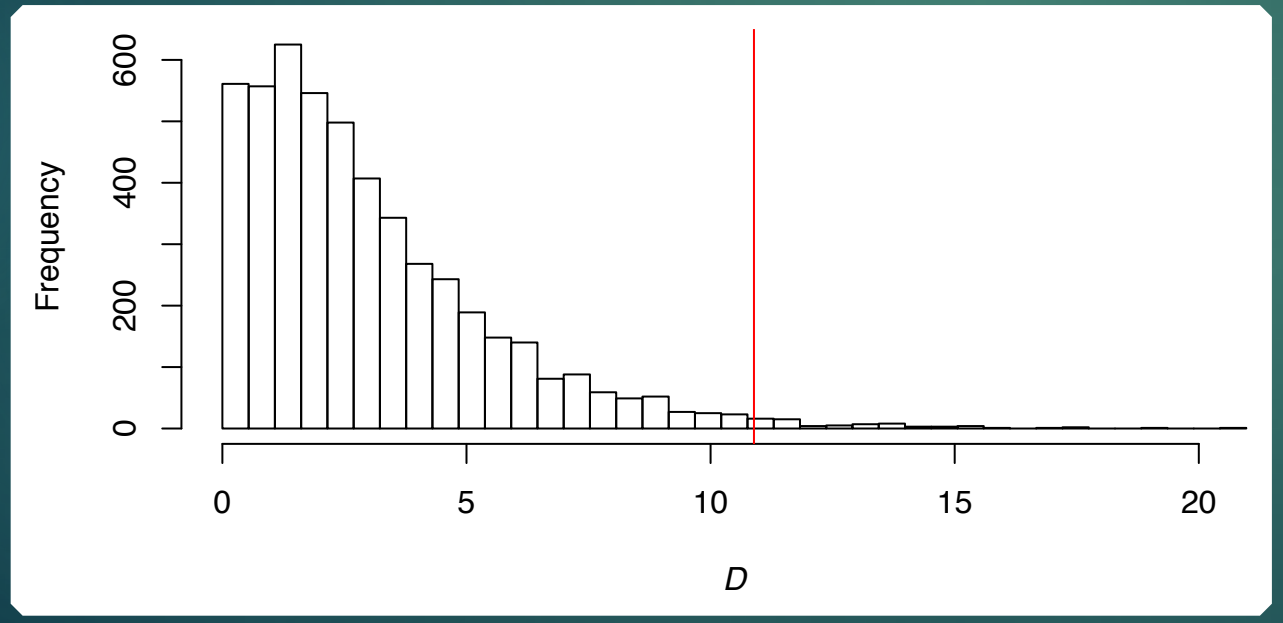
2. Replication Hypothesis H0

- Informative hypothesis: e.g., $\beta_1 > \beta_2$, $\beta_1 > 3.0$, $\beta_2 = 2.0$, etc.

```
parameter  plabel    inf  value
sb         ".p124." "<"  "-0.737"
qb         ".p125." ">"  "0.031"
t.pGender  ".p29."  ">"  "4.996"
t.guilt    ".p45."  ">"  "0.911"
t.anger    ".p37."  ">"  "1.771"
t.threat   ".p53."  ">"  "2.76"
sw.pGender ".p27."  "<"  "-0.217"
sw.anger   ".p25."  "<"  "-0.069"

H0 <- paste(hyp[,1],hyp[,2],hyp[,3],collapse="&")
```

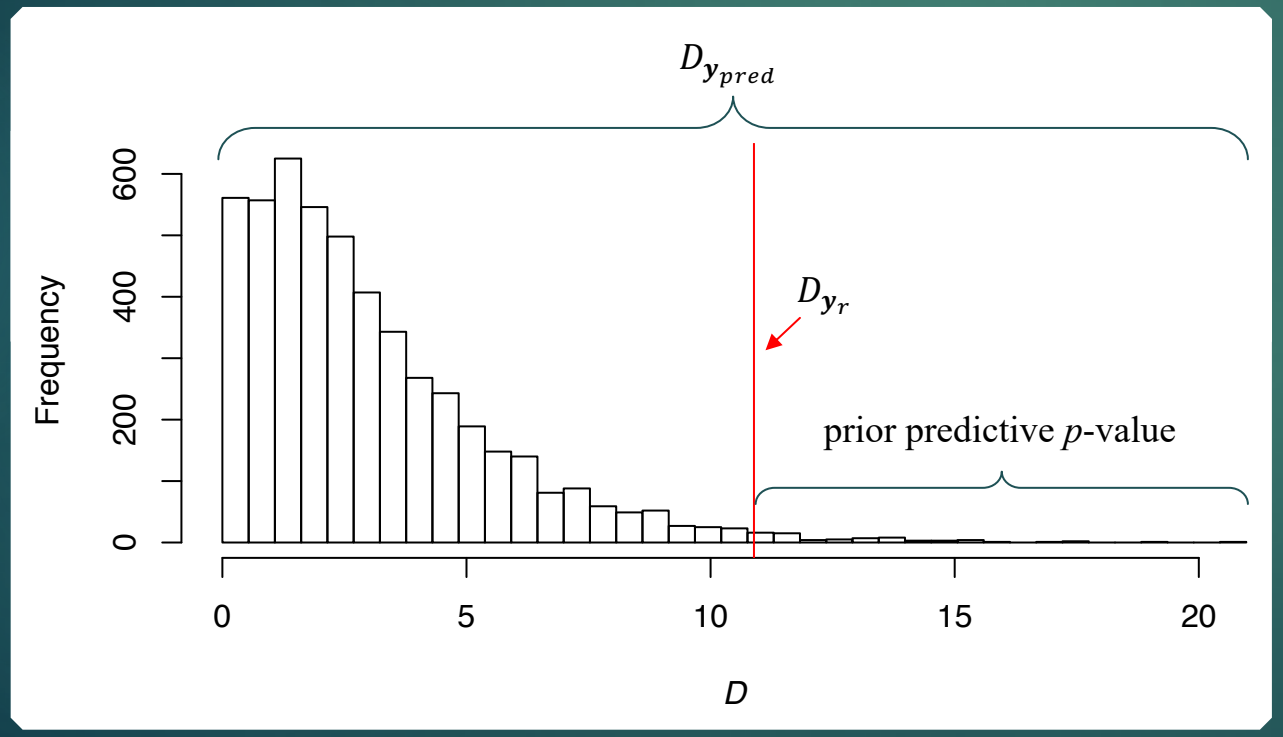
3. Prior predictive *p*-value

```
y.r <- read.table("yr.txt",header=TRUE) #new. data
step23.B <- ppc.step2step3(step1=step1,y.r=y.r,
                           model=model,H0=H0)
```



D_r = 10.89
Prior predictive *p* = .013

3. Prior predictive *p*-value



D_r = 10.89
Prior predictive *p* = .013

Summary & Discussion

- The replicability of SEM needs to be evaluated
- Vote-counting and model fit do not test replication
  (*Note*. model fit can theoretically be used as a statistic for the prior pred. *p*-value)
- The prior predictive *p*-value is suitable to test replication of study results

- Challenges: define H0
- See: Zondervan-Zwijnenburg, M.A.J. (2019). How to Test Replication for Structural Equation Models. *PsyArXiv*. doi: 10.31234/osf.io/uvh5s