MEASUREMENT AND THE NEXT GENERATION OF OPEN SCIENCE REFORM

Dr. JK Flake | University of British Columbia

The plan

1. Measurement and replicability
2. What is schmeasurement?
3. Challenges for Replication Research
4. BIG Team Science and Registered Reports
5. The Need for Open Science Methodological Development

A brief history

- B.S. in psychology with an area of concentration in statistics
- Master’s in quantitative psychology
- PhD in educational psychology from the measurement, evaluation, and assessment program
- Post doc in educational psychology and quantitative psychology
- Assistant to associate prof of quantitative psychology
- Assistant prof of quantitative psychology

NKU NORSE

JMU
JAMES MADISON
UNIVERSITY

UCONN
UNIVERSITY OF CONNECTICUT

YORK
UNIVERSITÉ
UNIVERSITY

UNIVERSITY of VIRGINIA

McGill

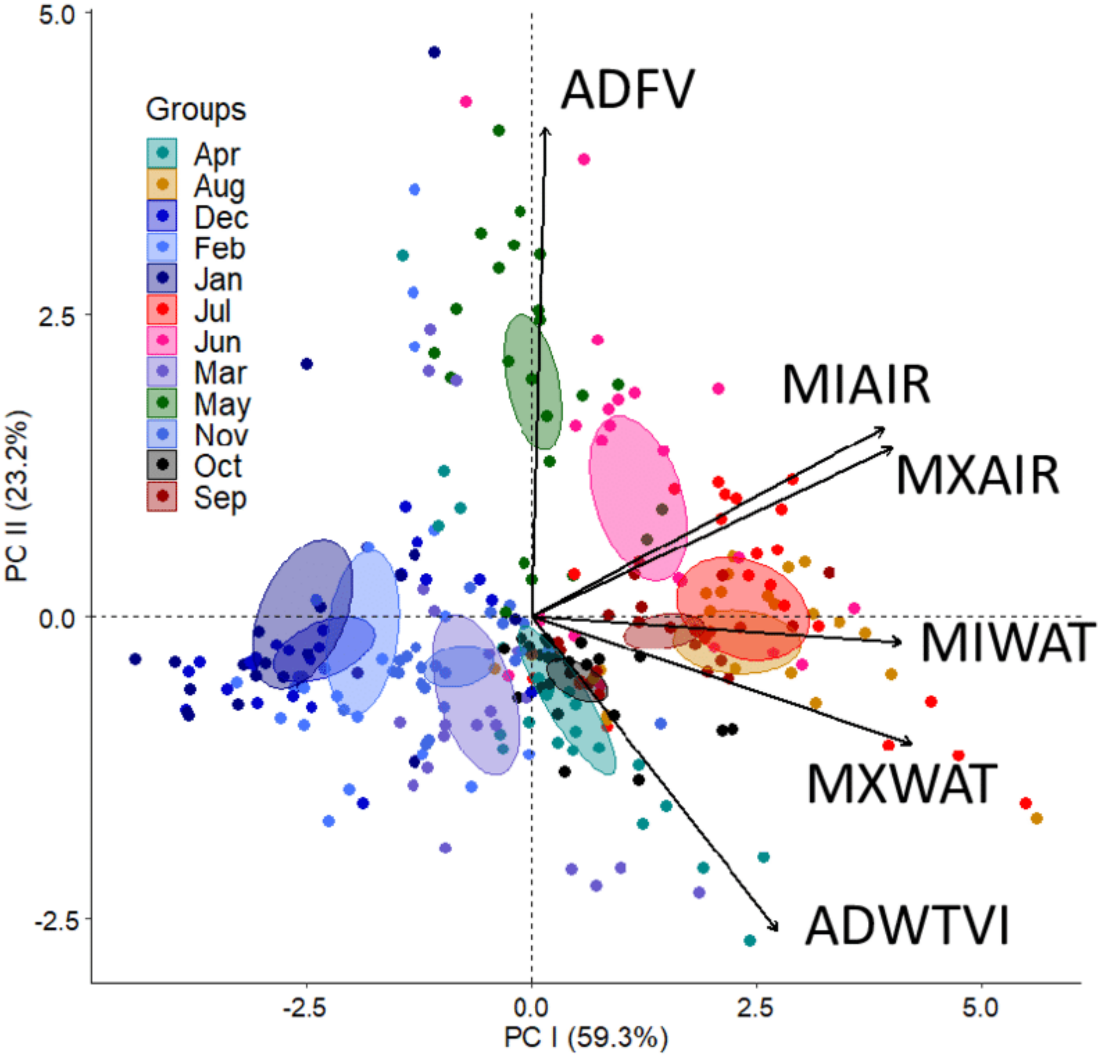
Extra Tooth History

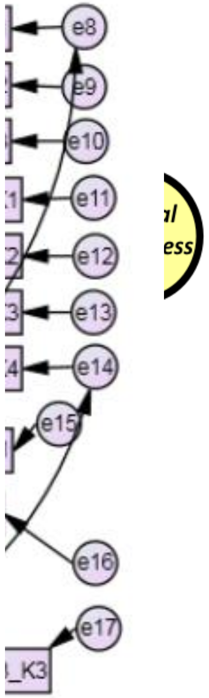
- This isn't the first time academic family have helped me in a bind!
- My first month of graduate school I was in bike accident
- I thought I might have to drop out of school, but my new academic family saved me
- Thanks to you all here and Tommaso, it was smooth sailing last night getting a temporary tooth

MEASUREMENT AND REPLICABILITY

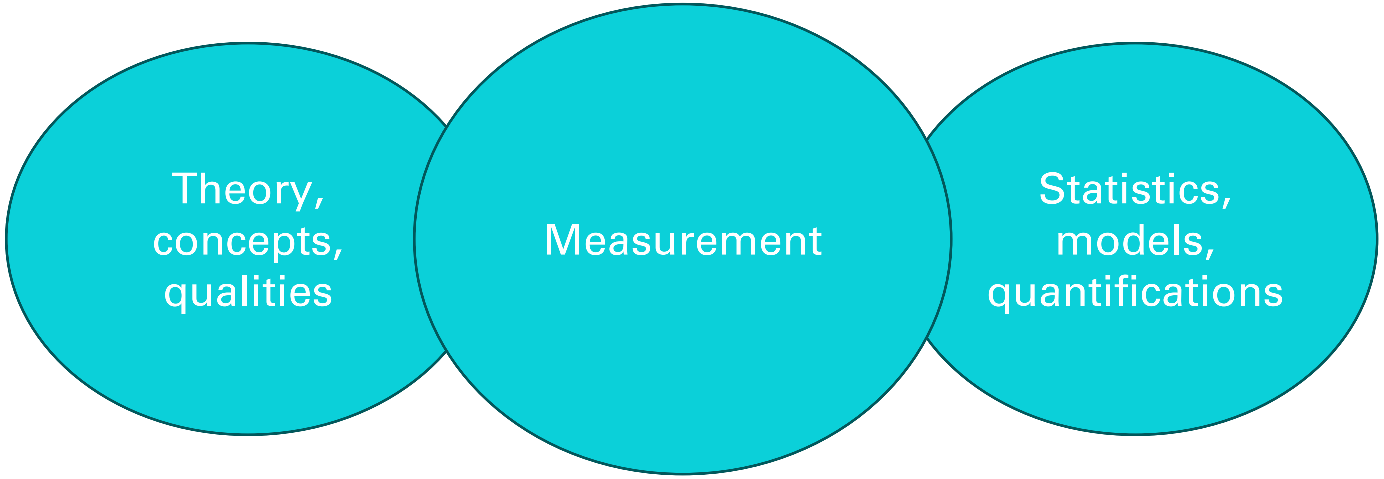
What do I mean by measurement and why does it matter for replication?

Psychological measurement

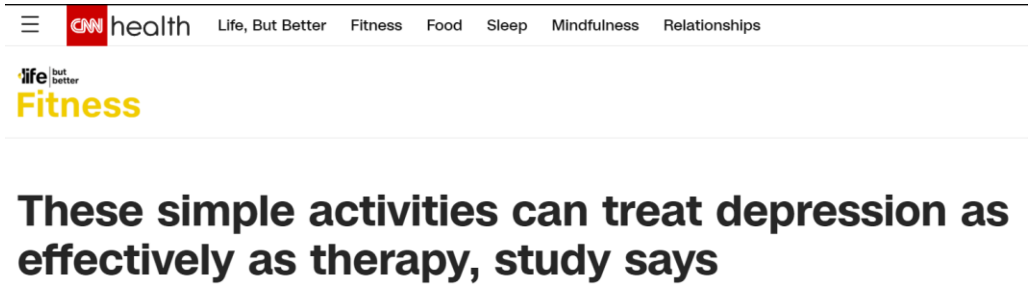




Complexity of measurement

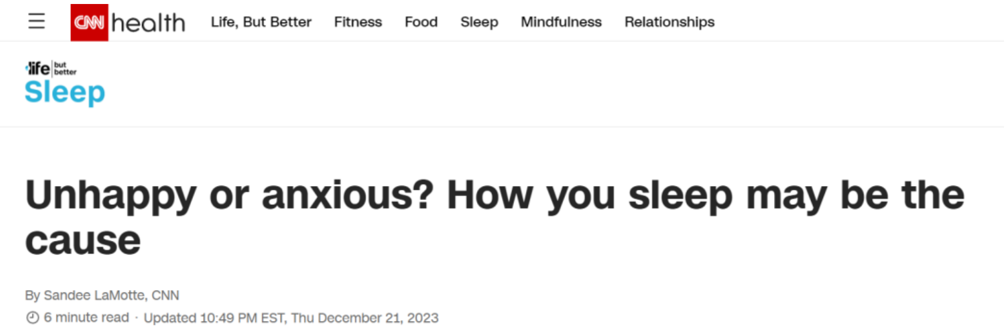

Theory, concepts, qualities
Measurement
Statistics, models, quantifications

Skeptical about any of these headlines?



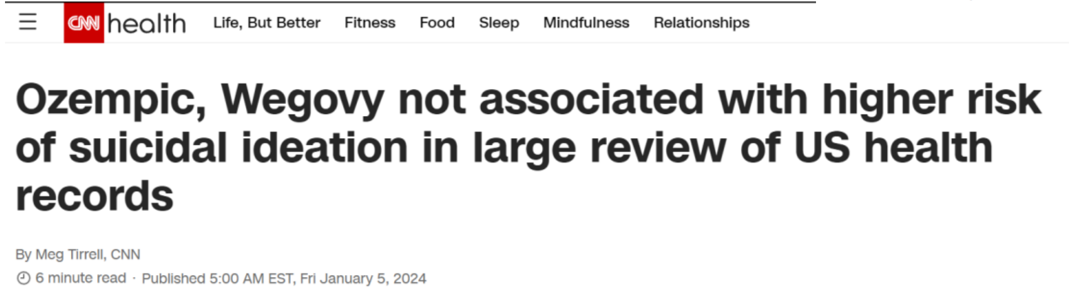
These simple activities can treat depression as effectively as therapy, study says

By Madeline Holcombe, CNN
3 minute read · Updated 7:14 PM EST, Wed February 14, 2024



Unhappy or anxious? How you sleep may be the cause

By Sandee LaMotte, CNN
6 minute read · Updated 10:49 PM EST, Thu December 21, 2023

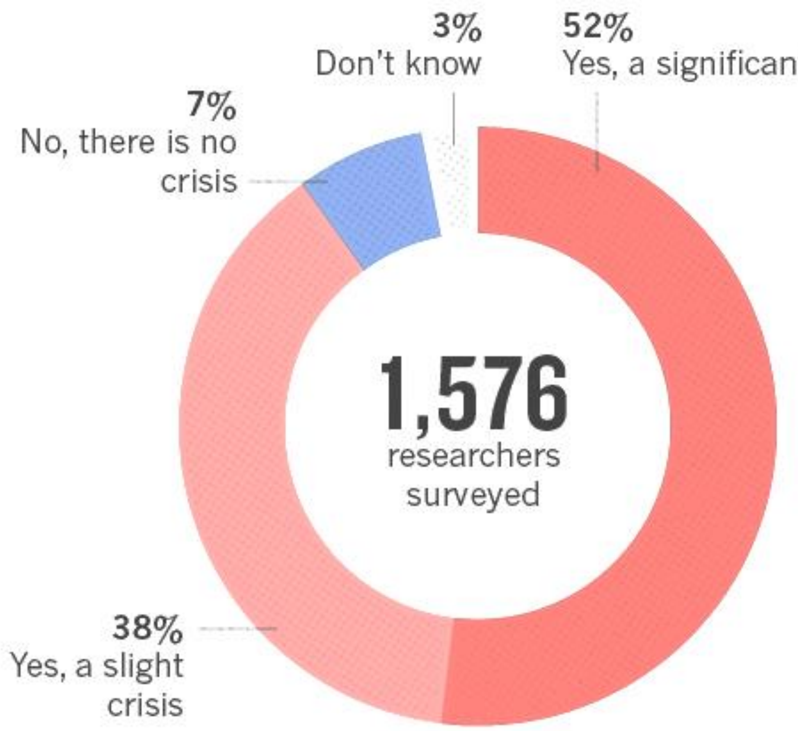


Ozempic, Wegovy not associated with higher risk of suicidal ideation in large review of US health records

By Meg Tirrell, CNN
6 minute read · Published 5:00 AM EST, Fri January 5, 2024

Crisis of confidence

IS THERE A REPRODUCIBILITY CRISIS?



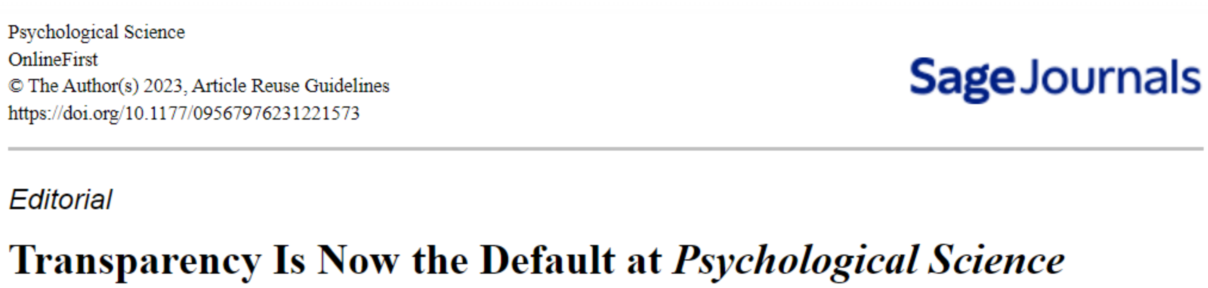
©nature

The open science movement

- Debates ensued, but if there was a war, it was won
  - Journals and funders started rolling out infrastructure and support
- Today, our government and leading journals endorse and pioneer in open science
  - Large scale replications become a norm
  - Registered reports take-off

ASPREDICTED

OSF



Psychological Science
OnlineFirst
© The Author(s) 2023, Article Reuse Guidelines
https://doi.org/10.1177/09567976231221573

Sage Journals

Editorial

Transparency Is Now the Default at *Psychological Science*

ROADMAP FOR OPEN SCIENCE

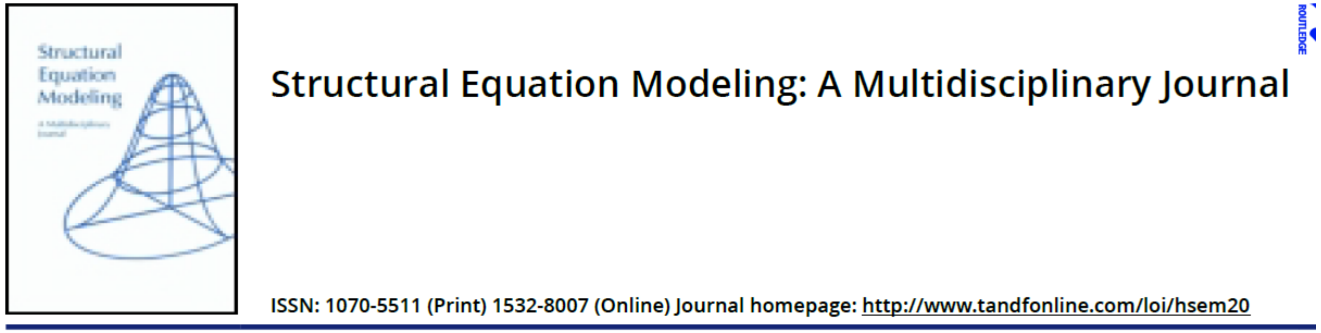
FEBRUARY 2020

Office of the Chief Science Advisor of Canada | Bureau du conseiller scientifique en chef du Canada

Canada

But why do I care?

- Back in 2015…
- Post-doc at York University
- Meta-science on measurement practices
  - We started with a couple dozen papers in JPSP
  - To date we've reviewed hundreds of original and replication studies' measures



Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: http://www.tandfonline.com/loi/hsem20

e Alignm
s Under
nce

ach

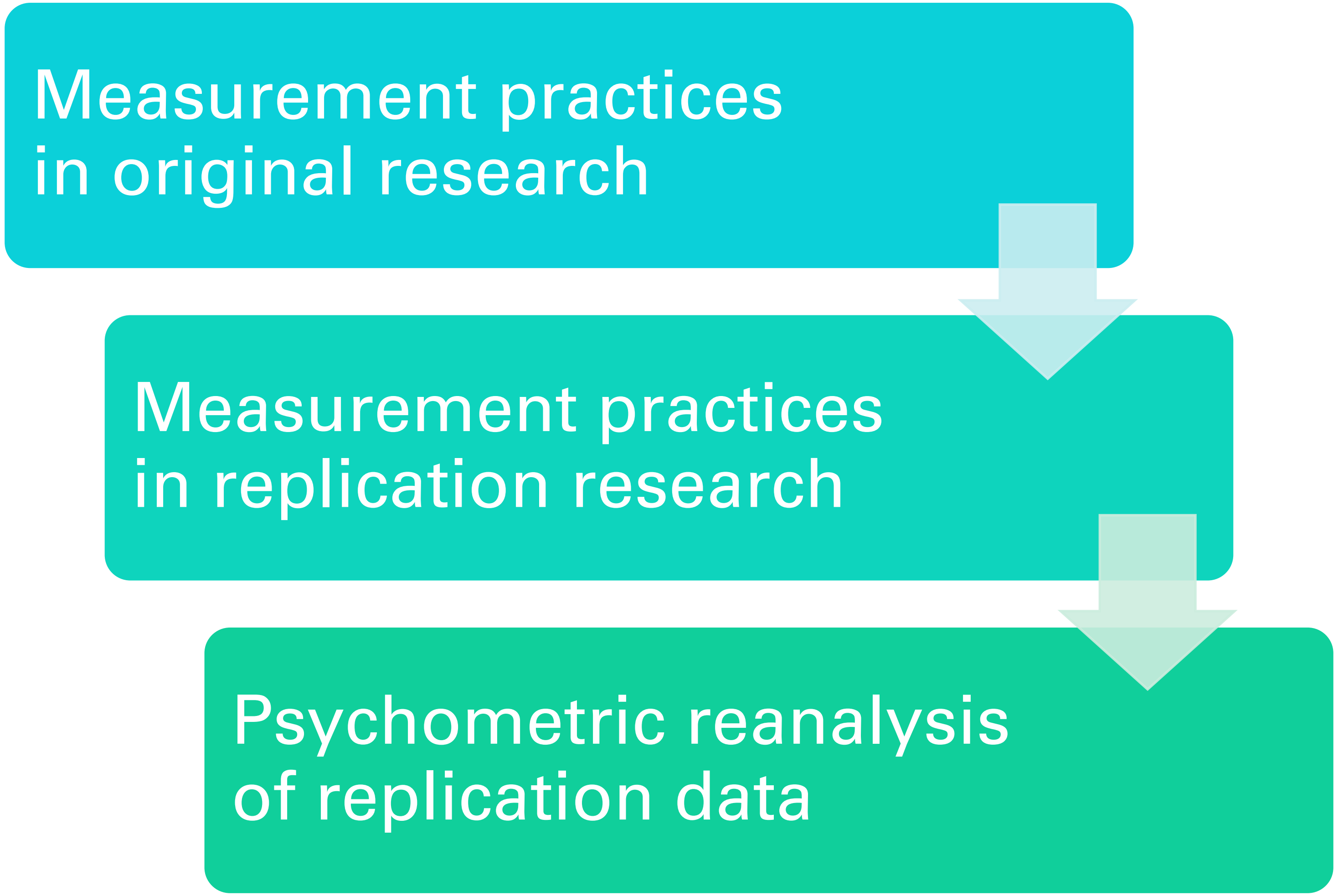
That all sounds complicated…

SCHMEASUREMENT

How are researchers using measures?

How is this related to replication?

Metascience journey



Measurement practices in original research

Measurement practices in replication research

Psychometric reanalysis of replication data

Review of measurement practices

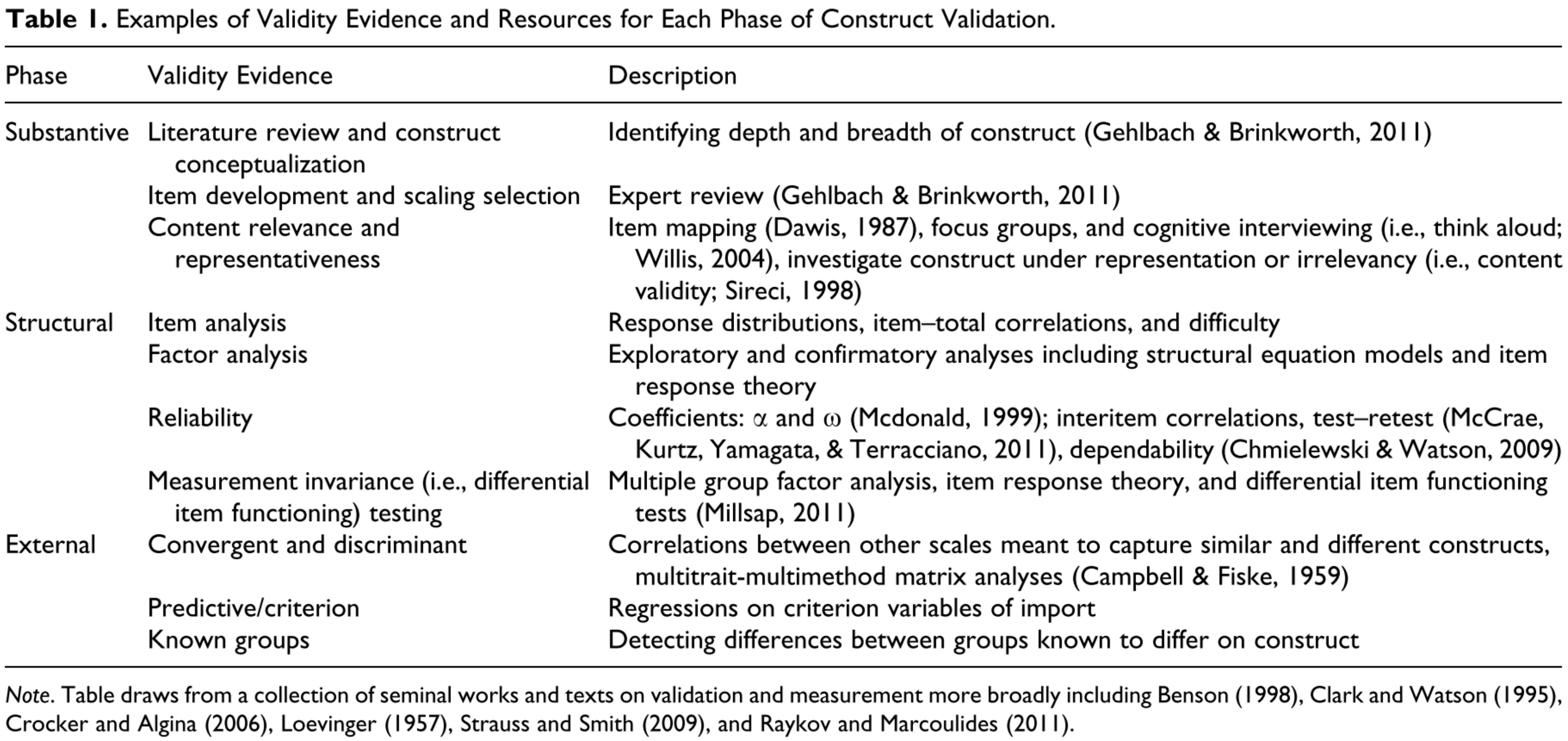


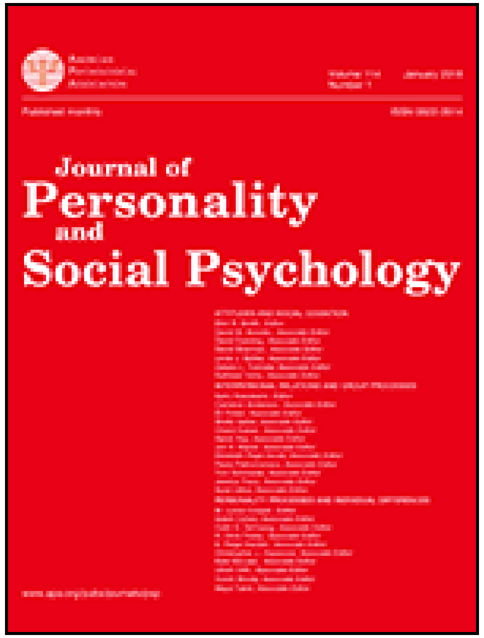
Table 1. Examples of Validity Evidence and Resources for Each Phase of Construct Validation.

Phase	Validity Evidence	Description
Substantive	Literature review and construct conceptualization	Identifying depth and breadth of construct (Gehlbach & Brinkworth, 2011)
	Item development and scaling selection	Expert review (Gehlbach & Brinkworth, 2011)
	Content relevance and representativeness	Item mapping (Dawis, 1987), focus groups, and cognitive interviewing (i.e., think aloud; Willis, 2004), investigate construct under representation or irrelevancy (i.e., content validity; Sireci, 1998)
Structural	Item analysis	Response distributions, item–total correlations, and difficulty
	Factor analysis	Exploratory and confirmatory analyses including structural equation models and item response theory
	Reliability	Coefficients: α and ω (Mcdonald, 1999); interitem correlations, test–retest (McCrae, Kurtz, Yamagata, & Terracciano, 2011), dependability (Chmielewski & Watson, 2009)
	Measurement invariance (i.e., differential item functioning) testing	Multiple group factor analysis, item response theory, and differential item functioning tests (Millsap, 2011)
External	Convergent and discriminant	Correlations between other scales meant to capture similar and different constructs, multitrait-multimethod matrix analyses (Campbell & Fiske, 1959)
	Predictive/criterion	Regressions on criterion variables of import
	Known groups	Detecting differences between groups known to differ on construct

Note. Table draws from a collection of seminal works and texts on validation and measurement more broadly including Benson (1998), Clark and Watson (1995), Crocker and Algina (2006), Loevinger (1957), Strauss and Smith (2009), and Raykov and Marcoulides (2011).

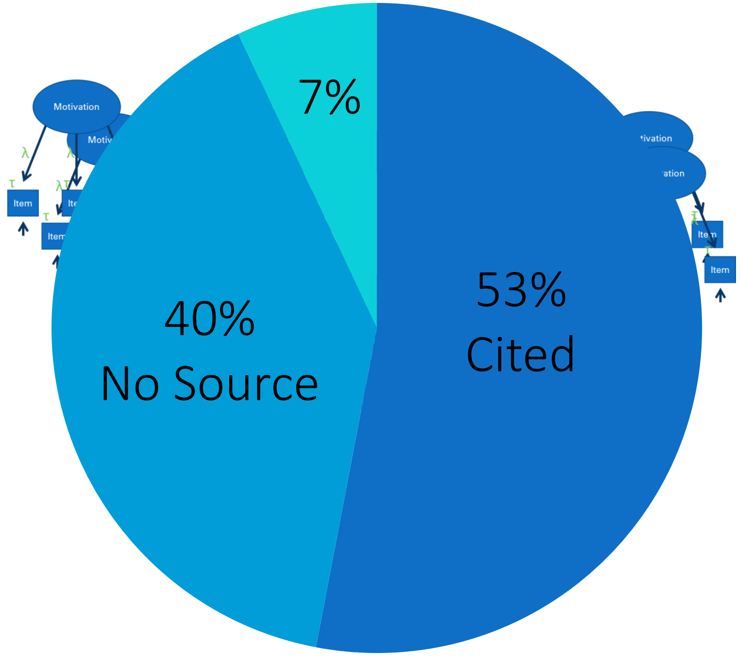
FLAKE, PEK & HEHMAN (*SOCIAL PSYCHOLOGICAL AND PERSONALITY SCIENCE,* 2017)

Review of measurement practices



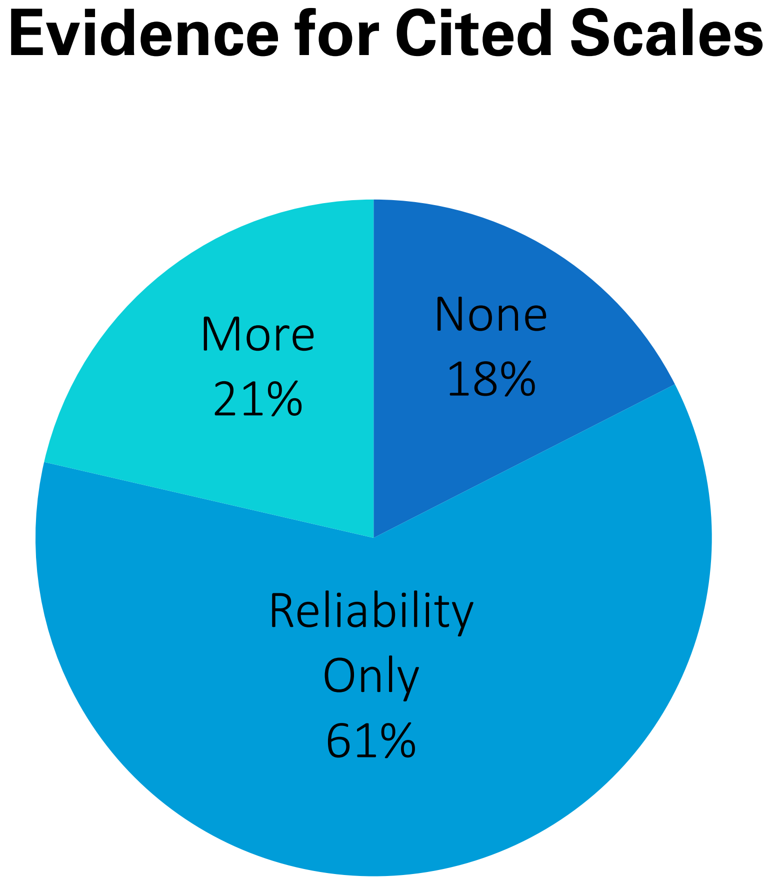
Journal of Personality and Social Psychology

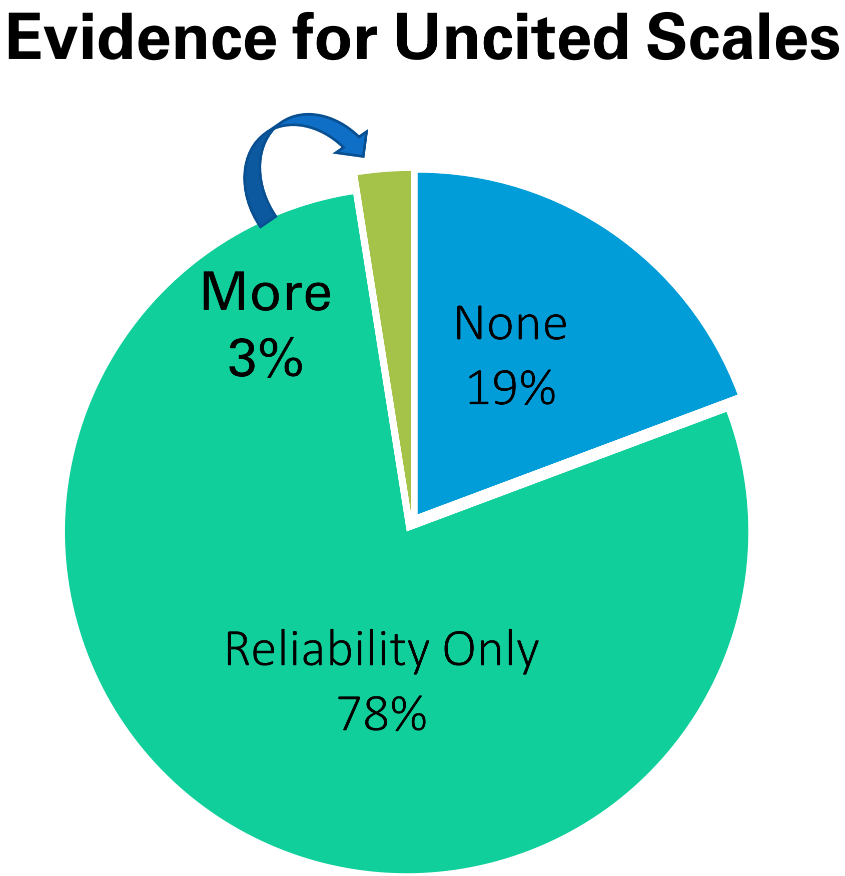
Coded 35 articles
700 instances of measures
87% were item-based scales
30% of those scales were 1-item



FLAKE, PEK & HEHMAN (*SOCIAL PSYCHOLOGICAL AND PERSONALITY SCIENCE,* 2017)

Reported evidence

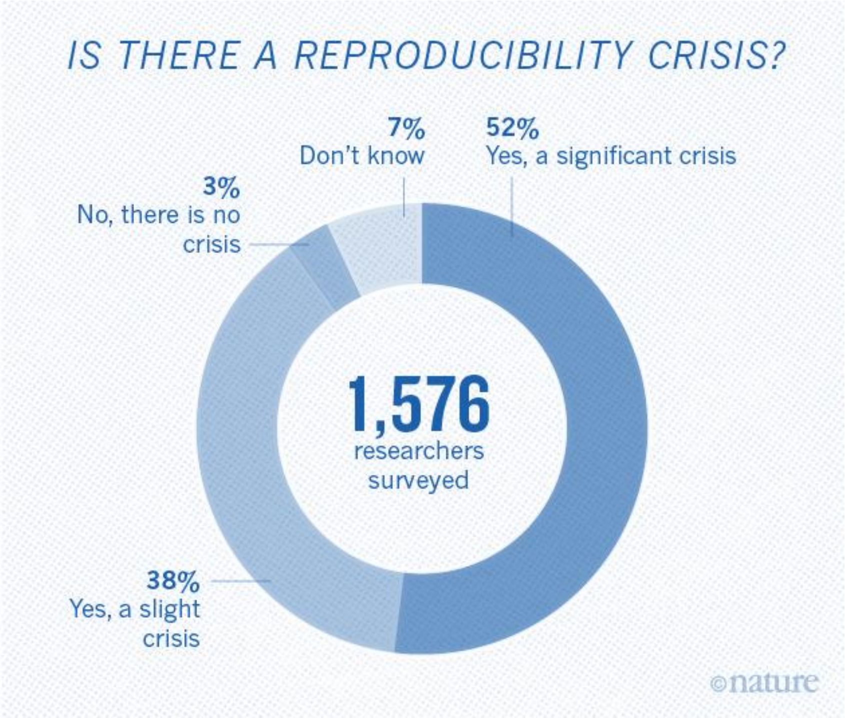


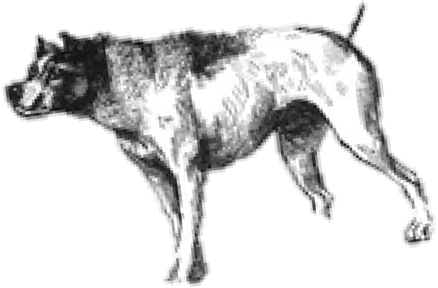


FLAKE, PEK & HEHMAN (*SOCIAL PSYCHOLOGICAL AND PERSONALITY SCIENCE,* 2017)

Connecting to the crisis

- Common unjustified (willy nilly) practices
    - Pulling apart or combining scales
    - Adding and removing items
    - Using scales made up on-the-fly
    - Using different sets of scales to measure the same thing across different studies
- Are these just ways to p-hack?





Questionable measurement practices (QMPs)

Questionable measurement practices are decisions researchers make that raise doubts about the validity of the measure use in a study, and ultimately the study's final conclusions

- QMPs raise doubts because of lacking justification and transparency
  - **Justification**: The reason for each specific decision
  - **Transparency**: Reporting of all decisions made, and how you made them, in the final work
- QMPs are not evidence of fraud or nefarious intent, they are just a lack of information
- QMPs make it difficult to impossible to evaluate the validity of the conclusion, and to reproduce and replicate studies

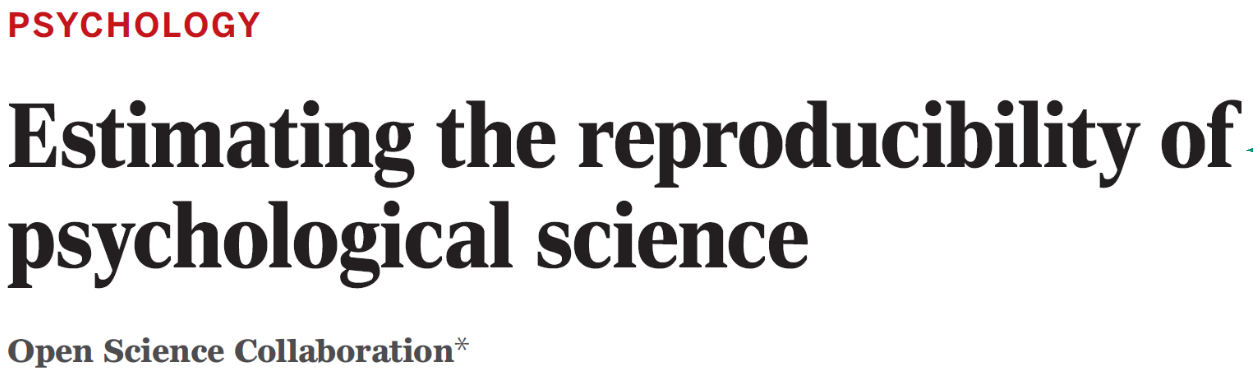
FLAKE & FRIED (*ADVANCES IN METHODS AND PRACTICES IN PSYCHOLOGICAL SCIENCE*, 2020)

CHALLENGES FOR REPLICATION RESEARCH

Are measures the same when you replicate?

What does the replication mean?

Measurement reviews of replication research



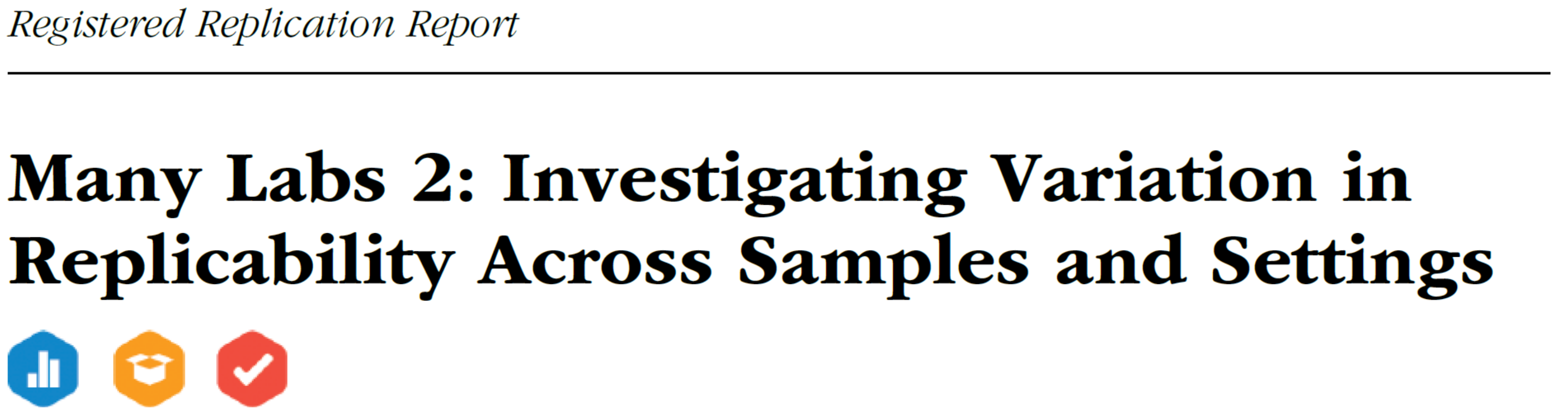
PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

1. 100 studies
2. Journal of Experimental Psychology: General, Journal of Personality and Social Psychology, and Psychological Science

1. Fewer studies, more data for each
2. Large open datasets for reanalysis



Registered Replication Report

Many Labs 2: Investigating Variation in Replicability Across Samples and Settings

Consistent results across reviews

- Original studies use hundreds of measures, mostly item-based scales
- Heavy reliance on…
  - single-item instruments
  - instruments made up on-the-fly
  - reliability
- ~20% of instruments reported with no information at all
- Replication studies report *even less* evidence
  - Less than half reported on reliability
- 16/100 of RPP reports explicitly indicated a measurement problem or concern
- It just wasn't common practice to evaluate the measures for these replication studies nor was (is) it clear how

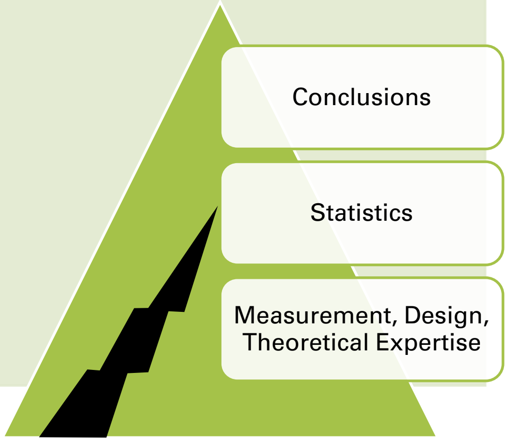
Measurement challenges for replication research

This is about score comparability. Measures that are comparable have similar statistical properties.

Limited information about measures

- #46, item wordings unclear, incorrect wordings ultimately used in replication study

No or limited validity evidence



Measurement differences

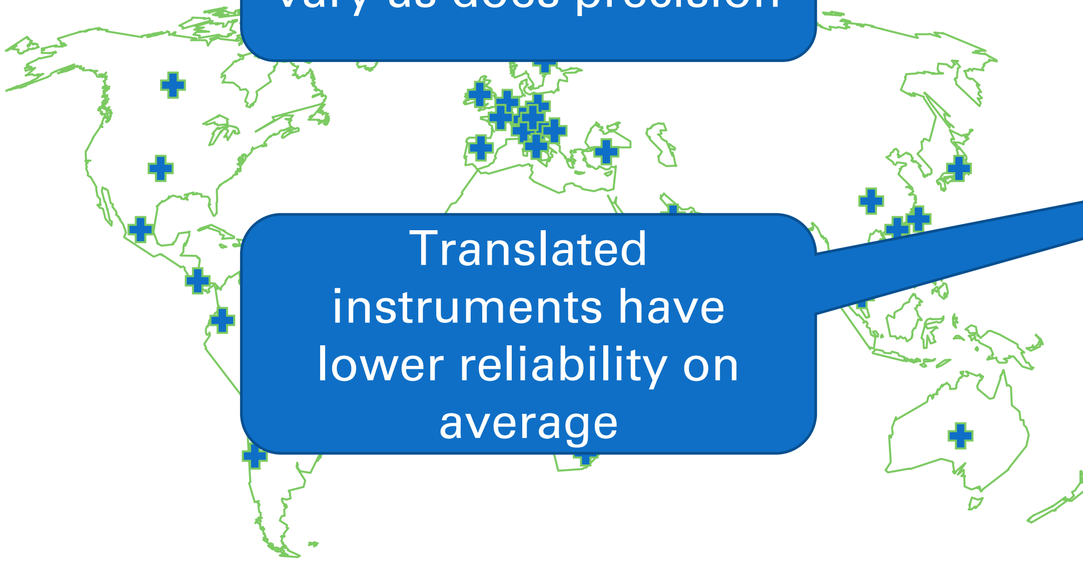
- #92 lower and unacceptable reliability
- #7 different number of factors

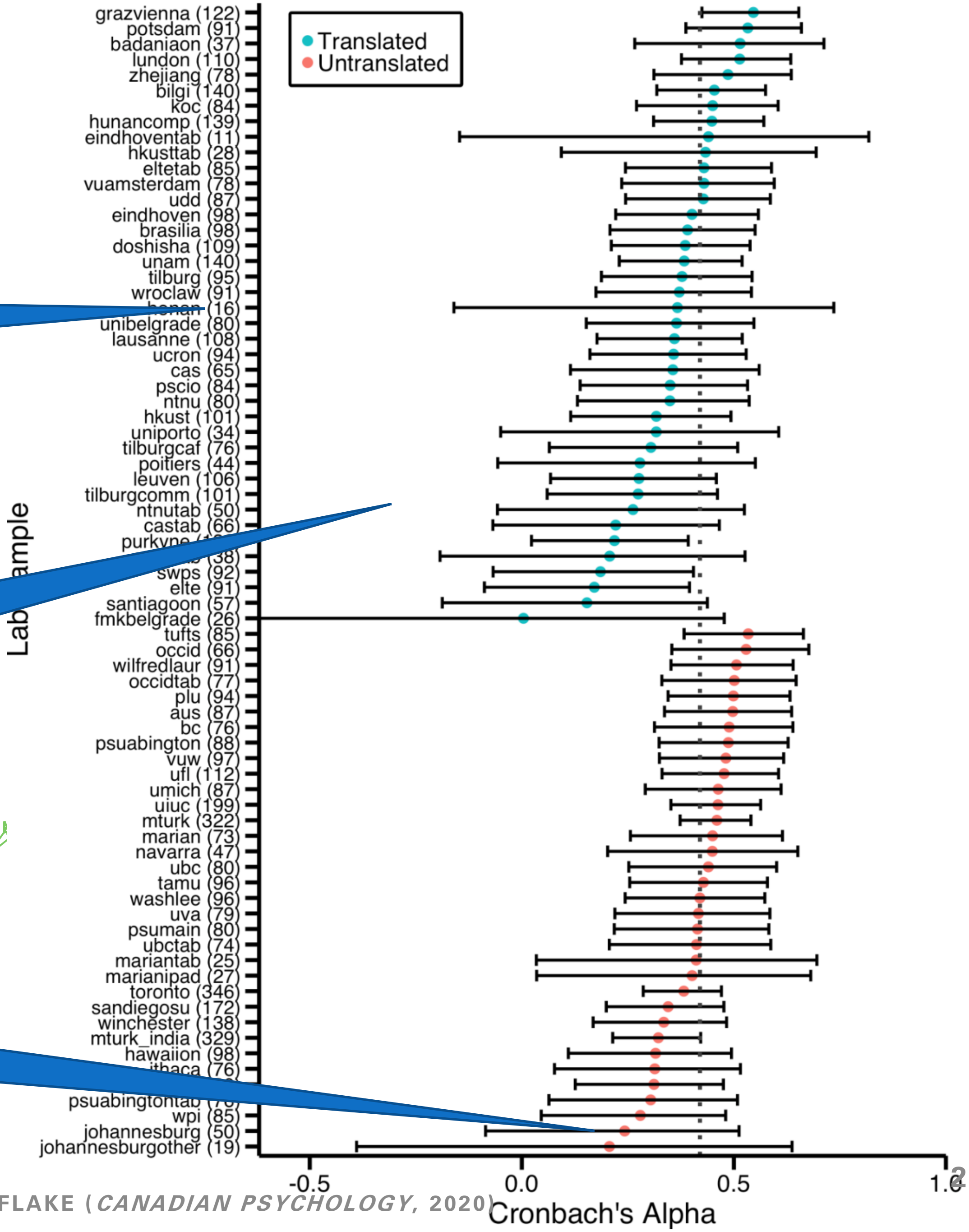
Translation

- 40 scales translated, only 8 were previously developed versions

FLAKE, DAVIDSON, WONG & PEK (*AMERICAN PSYCHOLOGIST*, 2022)

ML2 Measurement





Lab level sample sizes vary as does precision

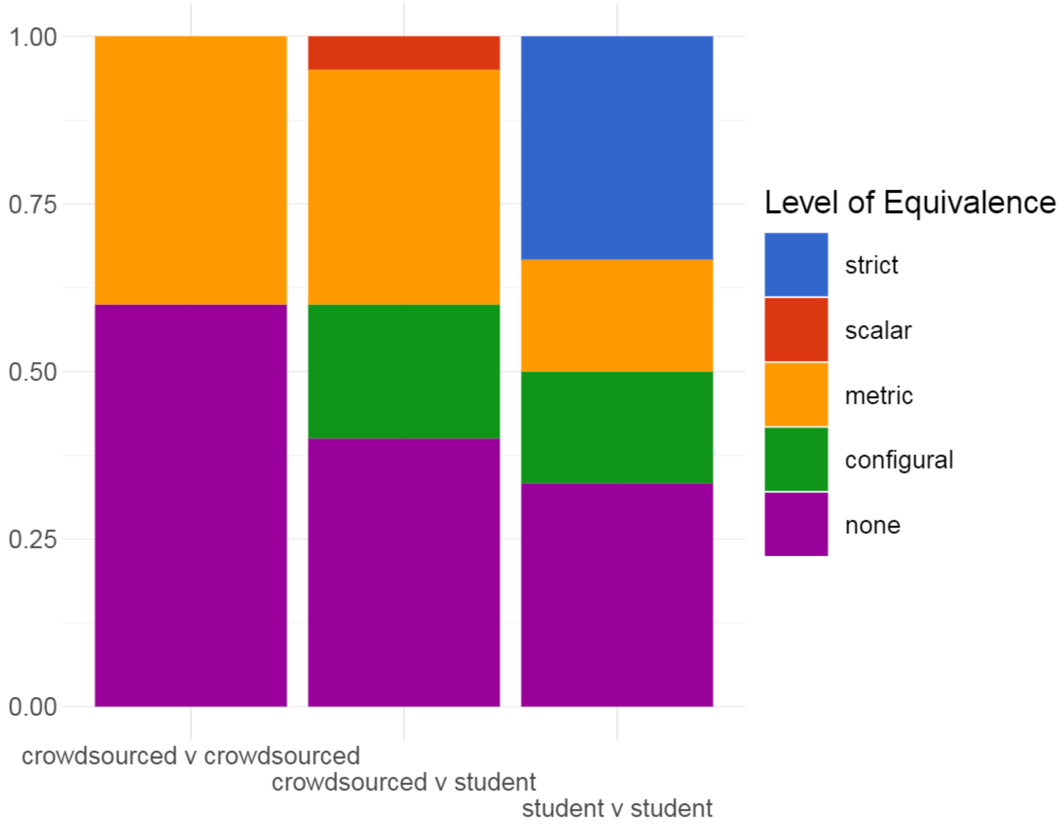
Translated instruments have lower reliability on average

Reliability is poor

*SHAW, *CLOOS, *LUONG, *ELBAZ & FLAKE (*CANADIAN PSYCHOLOGY*, 2020)

Evaluating equivalence

- Many Labs used a lot of single item and behavioral measures
- Of scales long enough, 40% could not be analyzed due to gross model misfit
- Of analyzed scales, many do not demonstrate equivalence across data collection modality
- Challenges here
  - Poor baseline psychometric properties
  - Small group level sample sizes



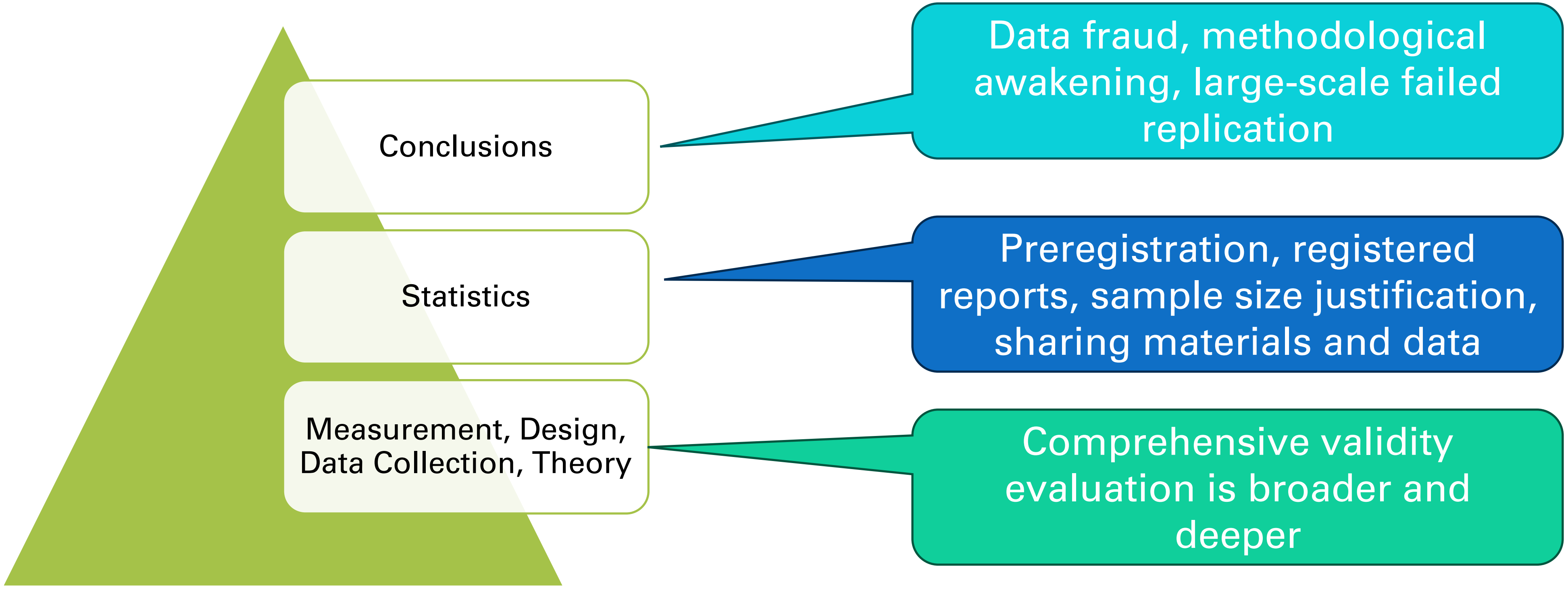
*ALLEY, AXT & FLAKE (*META PSYCHOLOGY*, IN PRESS)

Conclusions

- Measurement is not a trivial aspect of replication
- Lack of validity information and evidence can prevent replication
- Large scale replications introduce non-comparability
  - These are complex data structures with large potential for measurement heterogeneity
    - Translation, culture, and sampling methods are central to these studies
- If instruments do not produce comparable scores, results on combined data can be uninterpretable blends of population heterogeneity
- Sample sizes aren't large enough to evaluate instruments and their statistical properties
- Existing instruments tend not to be sound enough to evaluate

FLAKE (*EDUCATIONAL PSYCHOLOGIST*, 2021); FLAKE, *SHAW & *LUONG (*BEHAVIORAL AND BRAIN SCIENCES*, 2022)

The next generation of the methodological reform movement



Data fraud, methodological awakening, large-scale failed replication

Preregistration, registered reports, sample size justification, sharing materials and data

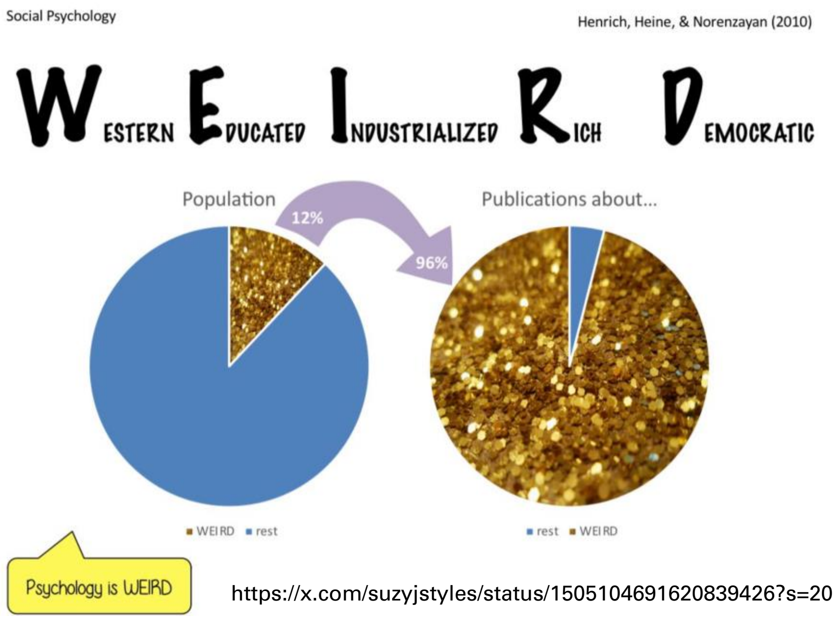
Comprehensive validity evaluation is broader and deeper

LARGE SCALE VALIDATION

How can we broaden methods reform to improve our research?

Methodological reform movement

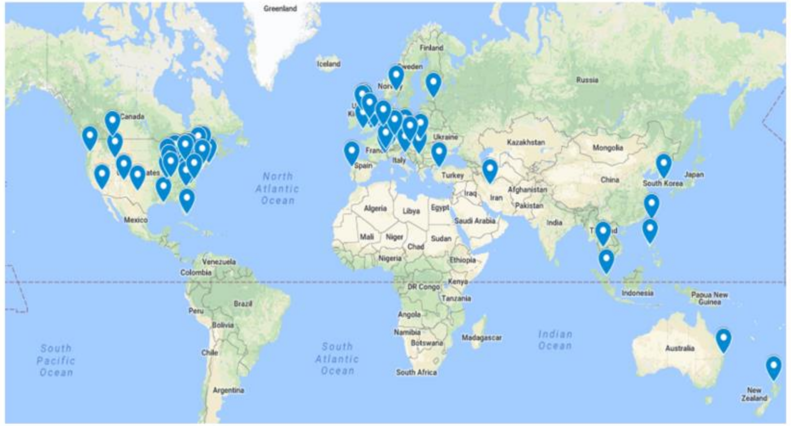
- Two (of many) major concerns born out of the replication crisis
  - Small and unrepresentative samples
  - Publication bias and the pressure to p-hack
- Two (of many) solutions that dovetail with concerns about representation in science
  - BIG team science
  - Registered Reports



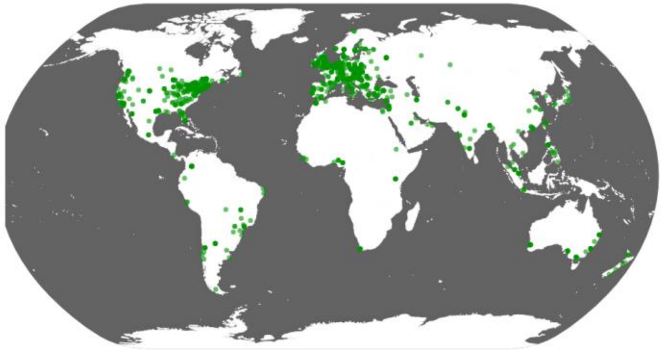
https://x.com/suzyjstyles/status/1505104691620839426?s=20

The Psychological Science Accelerator

- Chris Chartier pitched developing a CERN for psychology on his blog in 2017
- As a founding member I developed the data and methods committee
- Today we have over 3,000 members from over 80 countries
- Our first accepted study would go on to be one of the first registered reports published at *Nature: Human Behavior*



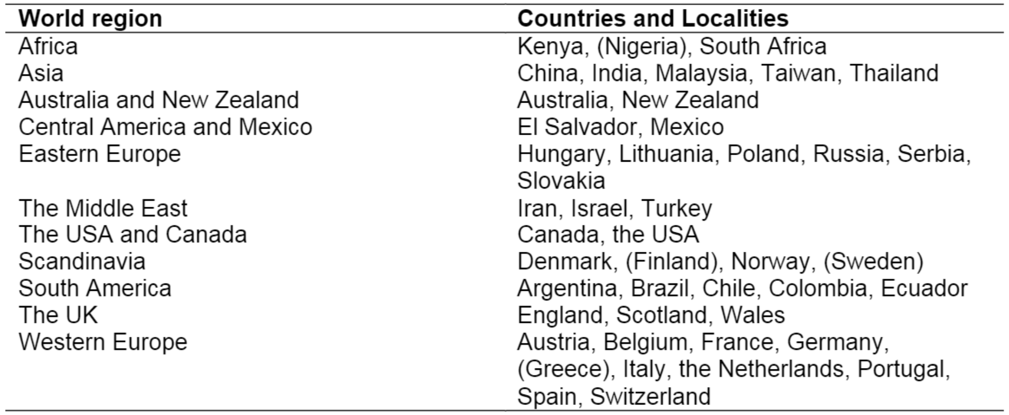
The Psychological Science Accelerator
1328 researchers, 84 countries



PSA001: Face perception around the world

- Large-scale replication of Oosterhof and Todorov’s valence-dominance model (2009)
- Data and stimuli are more diverse and representative than the original study
  - Over 11k participants
- Conducted by a diverse and international team of researchers
  - Over 100 authors
- Valence-dominance model varied across world regions
- Contributed to more nuanced and culture driven work in this area
- But, do the reusable materials and data have a bigger impact?

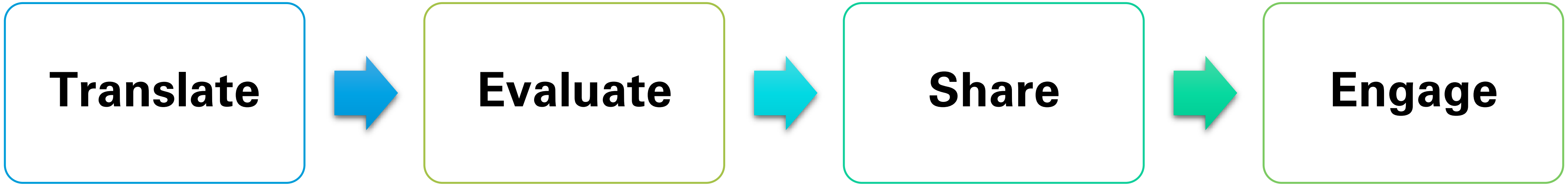
Photos from the Chicago Face Database used in PSA 001.



World region	Countries and Localities
Africa	Kenya, (Nigeria), South Africa
Asia	China, India, Malaysia, Taiwan, Thailand
Australia and New Zealand	Australia, New Zealand
Central America and Mexico	El Salvador, Mexico
Eastern Europe	Hungary, Lithuania, Poland, Russia, Serbia, Slovakia
The Middle East	Iran, Israel, Turkey
The USA and Canada	Canada, the USA
Scandinavia	Denmark, (Finland), Norway, (Sweden)
South America	Argentina, Brazil, Chile, Colombia, Ecuador
The UK	England, Scotland, Wales
Western Europe	Austria, Belgium, France, Germany, (Greece), Italy, the Netherlands, Portugal, Spain, Switzerland

JONES, DEBRUINE, FLAKE, ET AL. (*NATURE: HUMAN BEHAVIOR*, 2021)

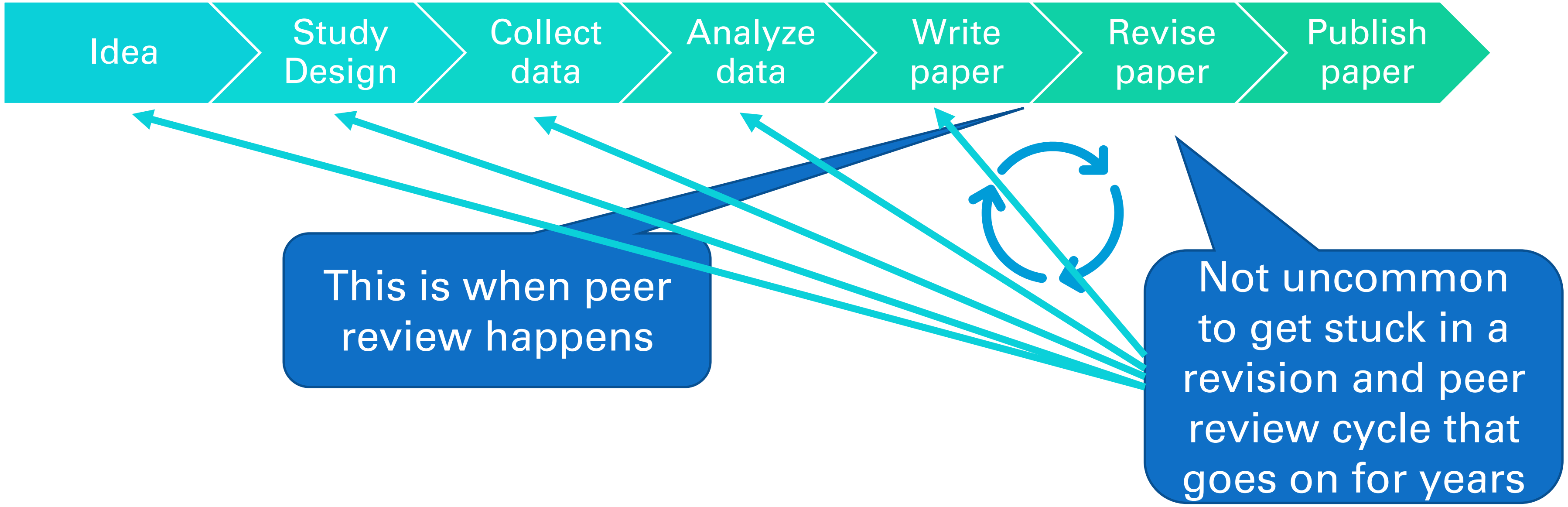
Large-scale validation



- Large scale replications have the potential to be engines for developing instruments that can be reused globally, if they are planned with measurement in mind
  - Collect enough data to evaluate the statistical properties of the instruments
- We used PSA 006 as a test case, the Oxford Utilitarianism Scale, translated into 23 languages
- Registered report to develop a measurement focused analysis pipeline that can be reused
  - Evaluate properties of instruments, assumptions of comparability, develop systems for transparent reporting and reproducibility

What is a registered report?

- First, how does the publication process for a manuscript typically work?

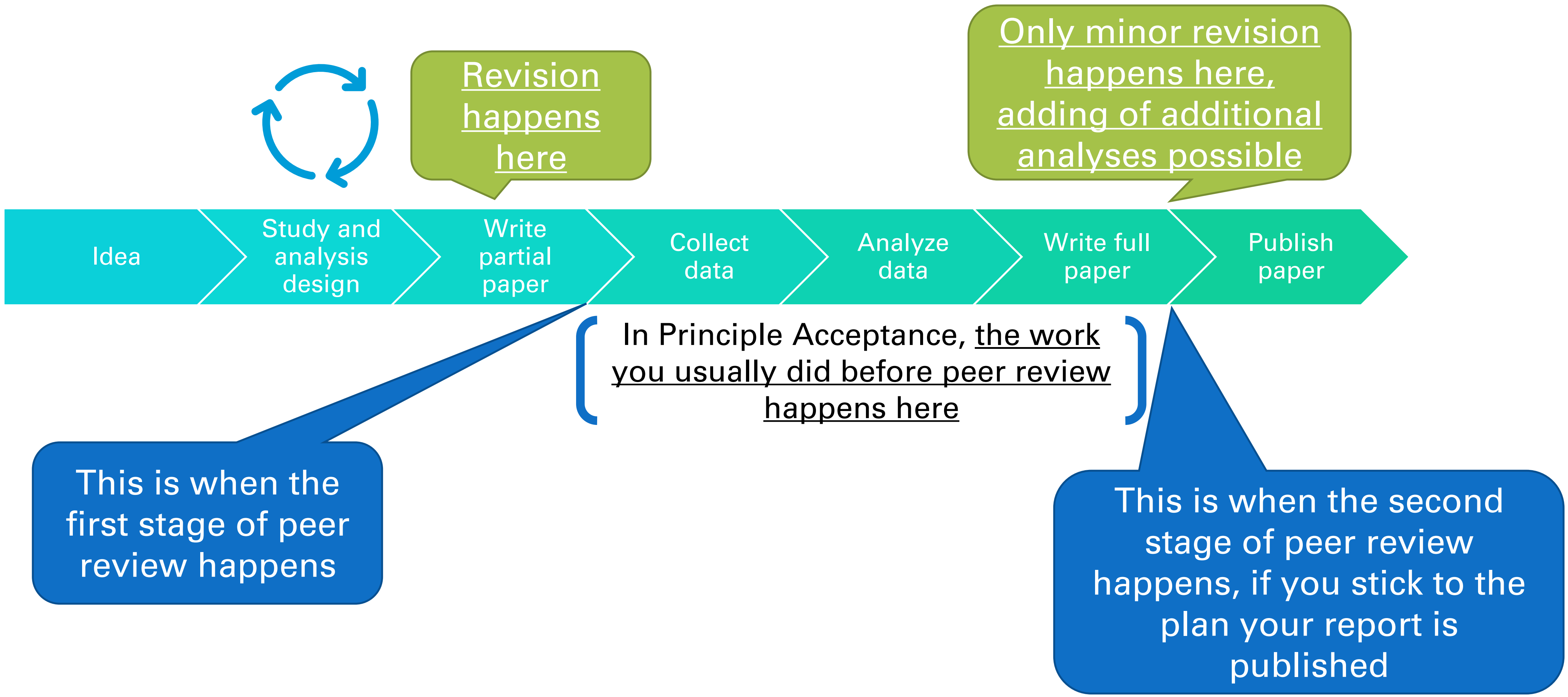


This is when peer review happens

Not uncommon to get stuck in a revision and peer review cycle that goes on for years

The registered report

Heard of preregistration? This uses the practice of preregistering, but instead of doing it on your own, you do it with a journal and it is peer reviewed



Idea → Study and analysis design → Write partial paper → Collect data → Analyze data → Write full paper → Publish paper

Revision happens here

Only minor revision happens here, adding of additional analyses possible

In Principle Acceptance, the work you usually did before peer review happens here

This is when the first stage of peer review happens

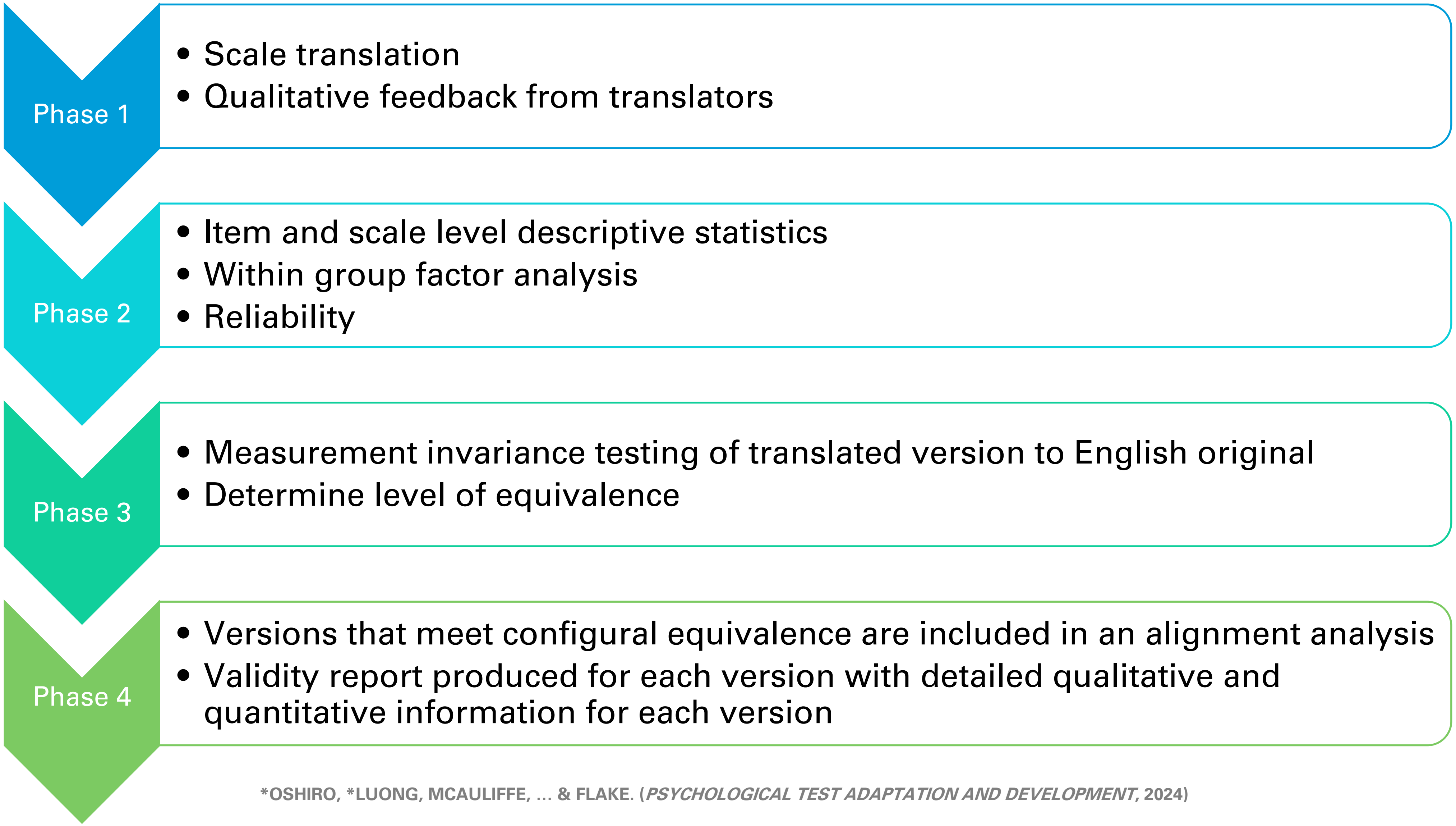
This is when the second stage of peer review happens, if you stick to the plan your report is published

Some pros of registered reports

- Limits publication bias
  - Results are not an aspect of the evaluation of the contribution of the paper
- Facilitates the design of studies that are useful regardless of outcome
- Facilitates reproducibility and replication
  - You can reproduce yourself
- Diversity and representation in science
  - Enables risky large-scale projects that otherwise wouldn't be feasible
    - Pooling resources across institutions
    - Multilab approaches to data collection with hard-to-reach populations
    - Less risk of being rejected because of controversial and/or unconventional results
- I'll leave the cons for discussion!

An RR is a good fit for the PSA because we can't afford to waste resources and need reusable code

Reproducible analysis pipeline



Phase 1

- Scale translation
- Qualitative feedback from translators

Phase 2

- Item and scale level descriptive statistics
- Within group factor analysis
- Reliability

Phase 3

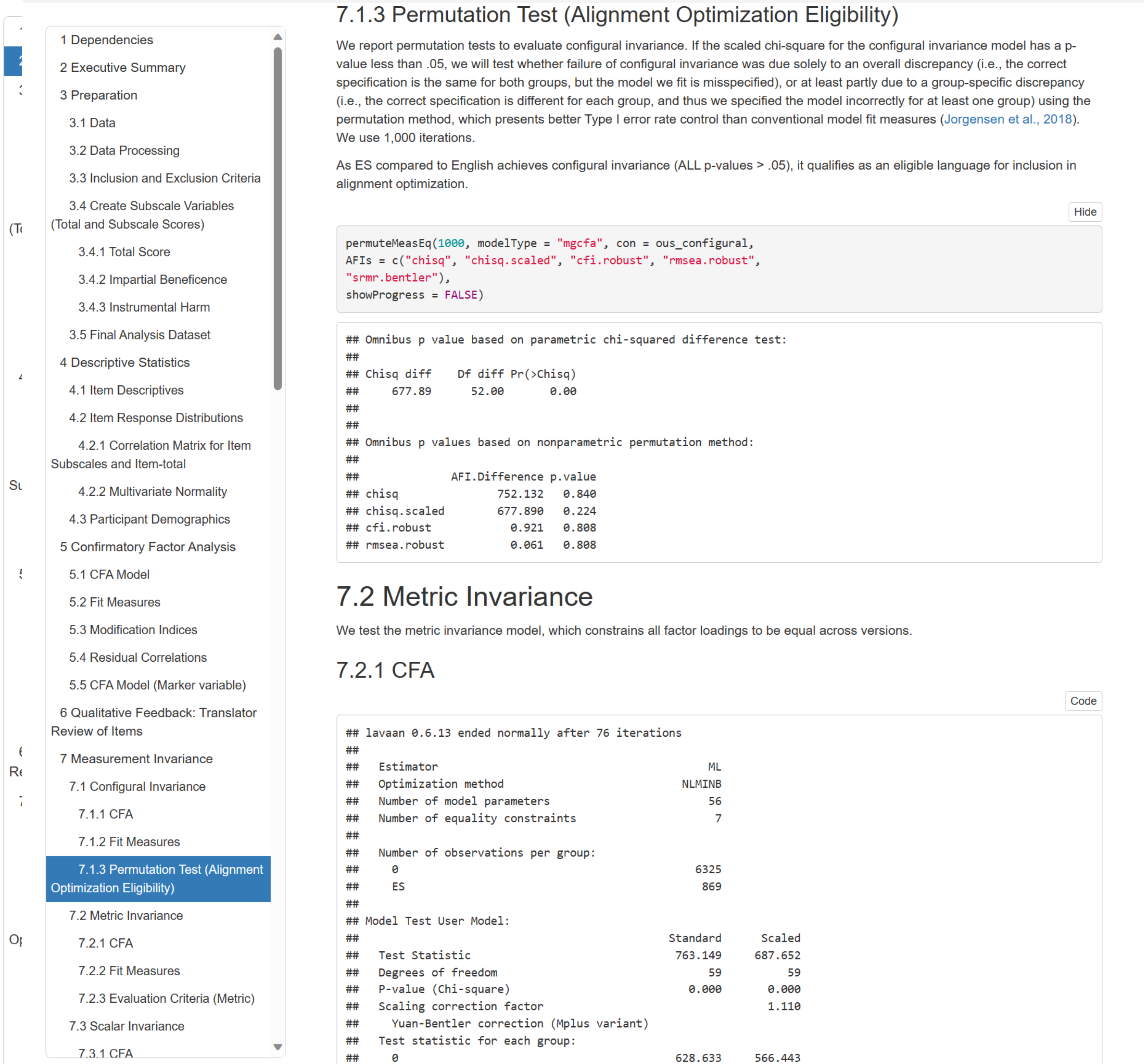
- Measurement invariance testing of translated version to English original
- Determine level of equivalence

Phase 4

- Versions that meet configural equivalence are included in an alignment analysis
- Validity report produced for each version with detailed qualitative and quantitative information for each version

*OSHIRO, *LUONG, MCAULIFFE, … & FLAKE. (*PSYCHOLOGICAL TEST ADAPTATION AND DEVELOPMENT*, 2024)

Ope



7.1.3 Permutation Test (Alignment Optimization Eligibility)

We report permutation tests to evaluate configural invariance. If the scaled chi-square for the configural invariance model has a p-value less than .05, we will test whether failure of configural invariance was due solely to an overall discrepancy (i.e., the correct specification is the same for both groups, but the model we fit is misspecified), or at least partly due to a group-specific discrepancy (i.e., the correct specification is different for each group, and thus we specified the model incorrectly for at least one group) using the permutation method, which presents better Type I error rate control than conventional model fit measures (Jorgensen et al., 2018). We use 1,000 iterations.

As ES compared to English achieves configural invariance (ALL p-values > .05), it qualifies as an eligible language for inclusion in alignment optimization.

```
permuteMeasEq(1000, modelType = "mgcfa", con = ous_configural,
AFIs = c("chisq", "chisq.scaled", "cfi.robust", "rmsea.robust",
"srmr.bentler"),
showProgress = FALSE)
```

```
## Omnibus p value based on parametric chi-squared difference test:
##
## Chisq diff    Df diff Pr(>Chisq)
##     677.89      52.00       0.00
##
##
## Omnibus p values based on nonparametric permutation method:
##
##              AFI.Difference p.value
## chisq               752.132   0.840
## chisq.scaled        677.890   0.224
## cfi.robust            0.921   0.808
## rmsea.robust          0.061   0.808
```

7.2 Metric Invariance

We test the metric invariance model, which constrains all factor loadings to be equal across versions.

7.2.1 CFA

```
## lavaan 0.6.13 ended normally after 76 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        56
##   Number of equality constraints                     7
##
##   Number of observations per group:
##     0                                             6325
##     ES                                             869
##
## Model Test User Model:
##                                               Standard      Scaled
##   Test Statistic                               763.149     687.652
##   Degrees of freedom                                59          59
##   P-value (Chi-square)                           0.000       0.000
##   Scaling correction factor                                  1.110
##     Yuan-Bentler correction (Mplus variant)
##   Test statistic for each group:
##     0                                          628.633     566.443
```

This is time consuming to develop, but it can increase the impact of your work. You don’t have to be a methods PhD to develop practices

Take-aways

- Psychometric properties are poor or mediocre for all versions
- Half of the instruments have a different configuration of items to factors than the original version
  - Not comparable at all
- More qualitative and conceptual research is needed to determine if there are cross-cultural differences in the construct (piu tardi)

My main take-away is that it was extremely difficult to register these psychometric analyses and I have a PhD in psychometrics. I'm also worried you might need a PhD in psychometrics to reproduce them…

METHODOLOGICAL DEVELOPMENT FOR OPEN SCIENCE

Where should we be headed next?

Let’s zoom in on just one aspect of that RR

- To use a multigroup factor model to compare two translated versions, you need to make a lot of decisions…

- Estimator
  - 7 options in Mplus
- Scale/model good enough to even try
  - 7: Model fit
    - **ONLY** commonly used (x2, RMSEA, SRMR, CFI, TLI, AIC, BIC) that could make thousands of combinations (7! = 5040)
  - 2: Reliability
    - **ONLY** commonly used
  - 2: Whole sample, groups separately
- Model identification (including anchor items)
  - Empirical methods
    - 2: Forward or backward
  - Evaluating significance
    - 4: LRT or 3 AFIs or some combination (4! = 24)
  - Pick one
    - 3: item review, lit review, 1st item
- Determination of levels
  - 8: Configural invariance
    - MGCFA model fit (at least 7 options and their combinations)
    - Permutation test
  - 10: Metric, Scalar, Strict
    - Model fit (at least 7 options)
    - Effect sizes (2 DMACS)

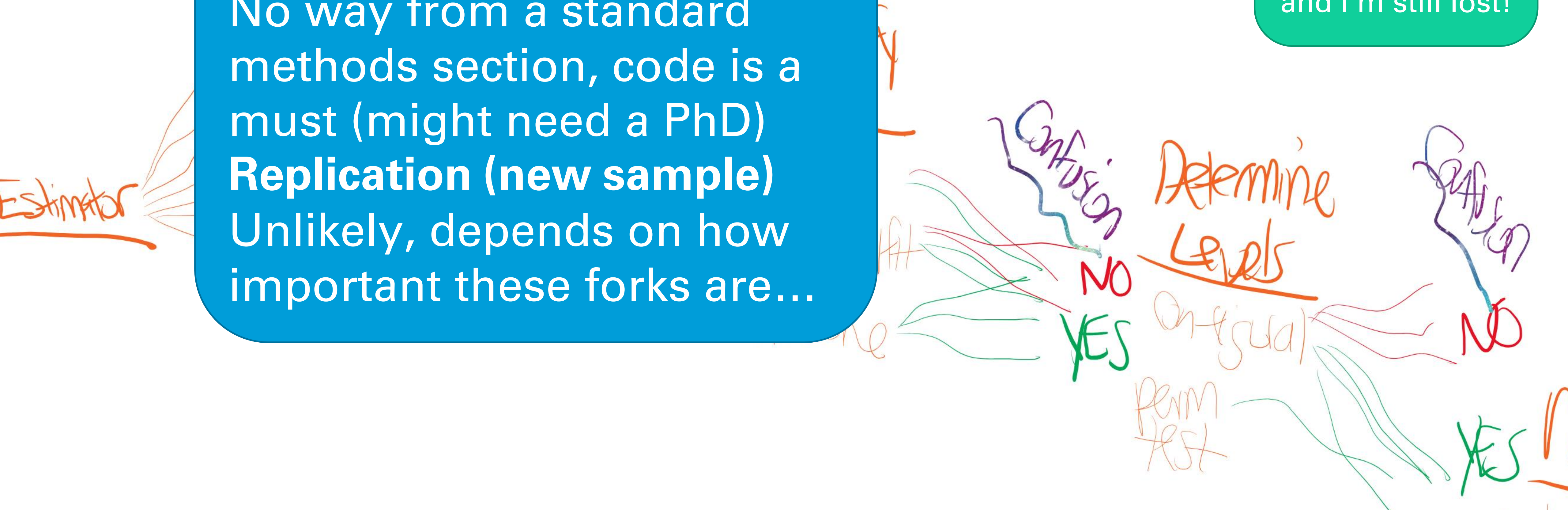
There are **at least** 45 decisions to navigate

Unhinged

Told you so…

I just read 10 simulation studies about this one decision and I'm still lost!

Reproducibility (same data)
No way from a standard methods section, code is a must (might need a PhD)
Replication (new sample)
Unlikely, depends on how important these forks are…



Methodological replication crisis

- Boulesteix discusses over optimistic bias in methods research
    - New methods are “better” than older ones and easier to publish
- More and more new methods means the garden of forking paths grows bigger and bigger
- Little incentive to publish accessible methods development
- Little incentive to publish methods work on how to navigate the garden

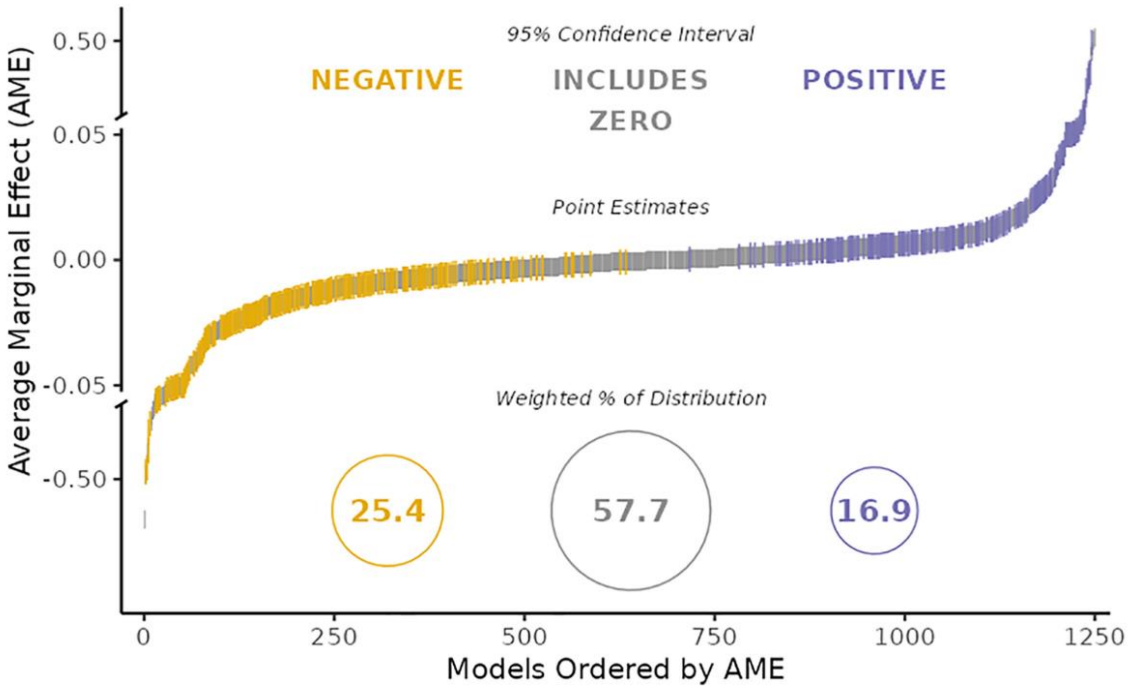
Between Researchers
Many possible reasonable paths through the garden, different people take different paths and get different results
Within Researchers
Wander through all the paths and get lost

Methodological research as cartography

- Heinze et al (2023) frames the problem as not fully developing methodologies
- Latter stages of development should neutrally compare methodologies in a wide array of real data situations and culminate in syntheses and reviews to develop practice
- We need to map the garden of forking paths
- We need methodological research that determines best practices for registered and transparent exploratory research

Some of what I’m thinking about now

- Integrating multiverse methods into simulation
  - i.e., many analysts and sensitivity analysis all seek to evaluate how robust a result is (e.g., Breznau, 2022 *PNAS*)
  - But they often use real data where the truth is unknown…
  - We could use this to develop maps of gardens of forking paths to enable registration
- Pushing methodologists to get their work to reproducible standards
- Learning how to register methodological and methodoligcally complex applied research
  - So I can help develop practices!



Active work in the lab

Mairead Shaw

Mapping the Multilevel Multiverse: Use simulation to see how sensitive results are to different paths that can be taken on the same data

Lindsay Alley

Coping with Baseline Model Misfit: Developing decision making criteria for researchers evaluating multiple group measurement models

Jacob Plantz

DIF Detection under Misspecification: Do results depend on forward or backward analysis paths?

Oulu Li-Tan

The Garden of Forking Structural Equation Modeling Paths: Systematic review of methods literature to develop multiverse protocols for real data

Take aways

- Measurement and modeling have a foundational role in replicability and reproducibility
  - Schmeasurement makes replication research difficult to conduct and the results uninterpretable
  - The methods and practices for this don’t exist yet
- Methodologies need to be developed to enable open science practices
  - Replicable measurement as a prerequisite to replication
  - Estimating and understanding result heterogeneity in relation to measures
  - Navigating methodological decisions and analytical flexibility
  - Transparency and reporting practices that can accommodate the complexity

THANK YOU MUCH! MERCI BEAUCOUP! GRAZIE MILLE!

Looking forward to questions and discussion