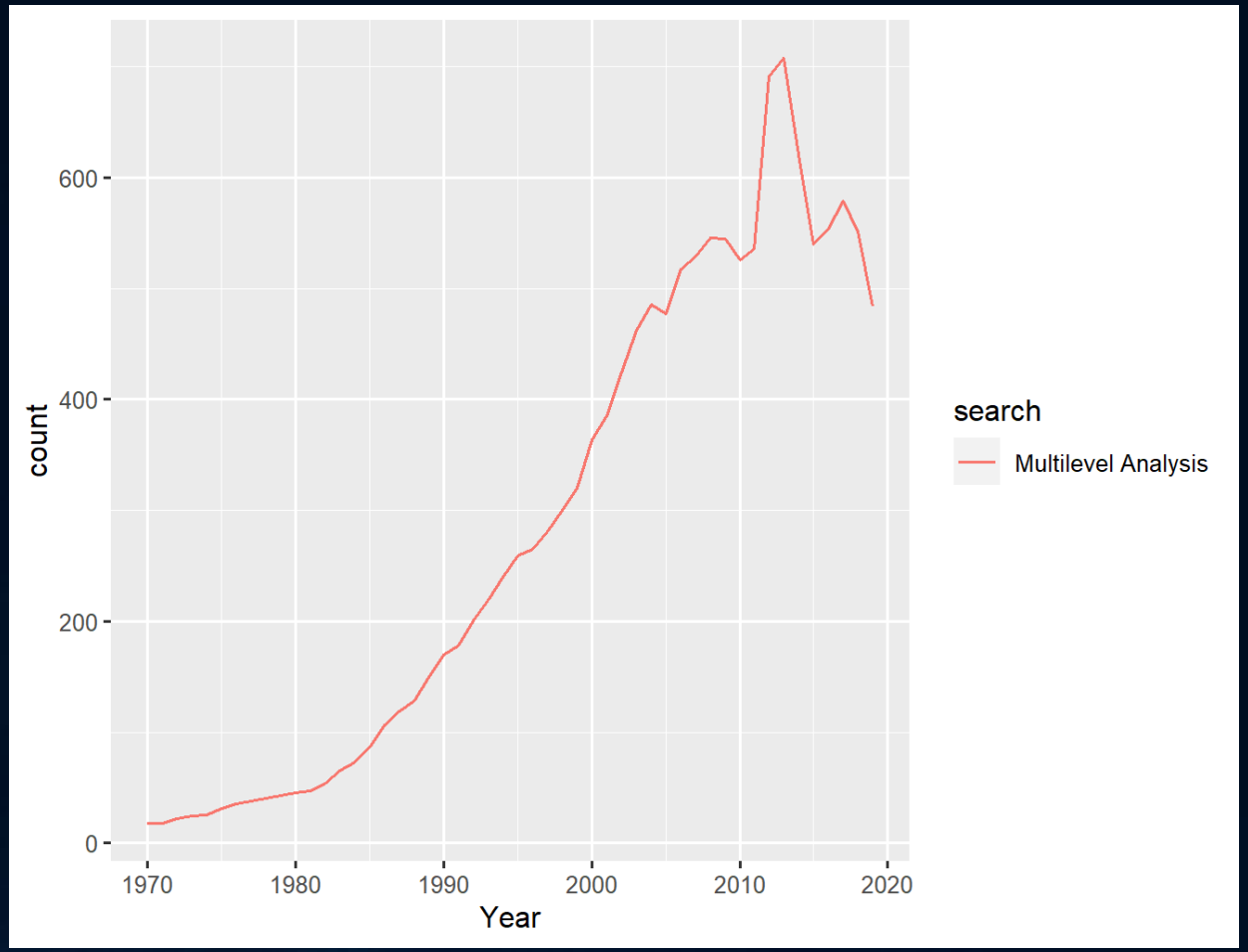
Data Aggregation in Multilevel Research: A Review & Recommendations for Improved Clarity, Transparency, & Reproducibility

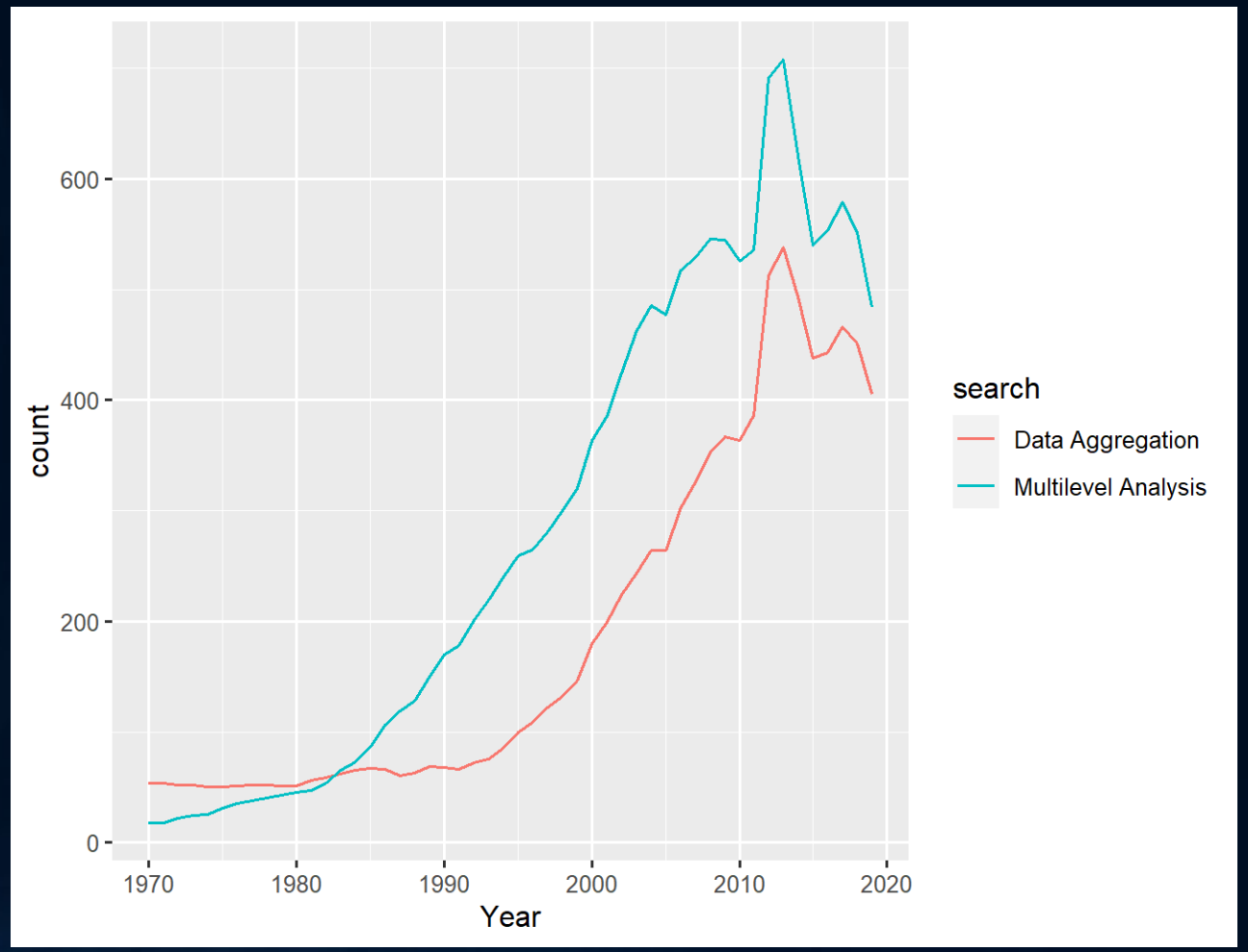
James M. LeBreton, Pennsylvania State University

with important contributions from
Amanda N. Moeller, Pennsylvania State University
Jenell L. S. Wittmer, University of Toledo

Growth of Multilevel Research



Growth of Multilevel Research



Advances in Multilevel Theory

- Meso Paradigm (House, Rousseau, & Thomas-Hunt, 1995)
- Composition Models and Multilevel Theory (Chan, 1998)
- Principles for Developing Multilevel Theory (Kozlowski & Klein, 2000)
- Bracketing to Expand Multilevel Research (Hackman, 2003)
- Homologous Multilevel Theories (Chen et al., 2005)
- Emergence as Process vs. Emerged Phenomena (Kozlowski et al., 2013)
- Integrating Dynamics and Change (Cronin & Vancouver, 2019)
- Expanded Perspectives on Emergence (Mathieu & Luciano, 2019)

Advances in Multilevel Measurement

- Understanding and Estimating Non-Independence
  - Kenny & Judd (1986); Bliese (2000); Bliese & Hanges (2004); Aguinis & Culpepper (2015)
- Estimating Within-Group Agreement
  - James et al. (1984; 1993); Lindell & Brand (1997); Burke et al. (1997); Brown & Hauenstein (2005); LeBreton et al. (2005); LeBreton & Senter (2008); Krasikova & LeBreton (2019); Newman & Sin (2020)
- Scaling & Centering
  - Hofmann & Gavin (1998); Enders & Tofighi (2007)

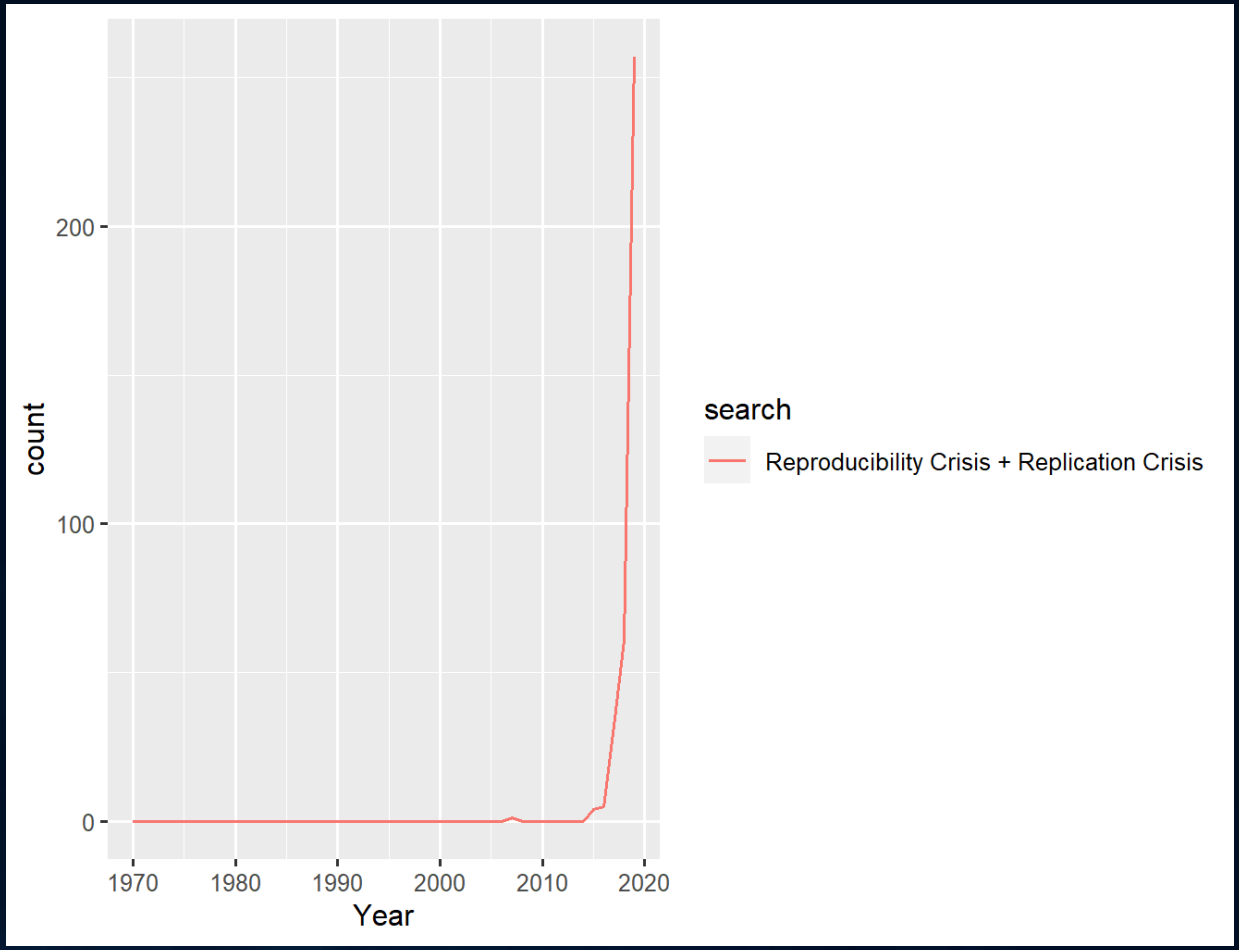
Advances in Multilevel Design

- Data Collection & Sampling
  - Beal & Weiss (2003); Beal & Gabriel (2019); Zhou et al. (2019)
- Power Analyses for Multilevel Inferences
  - Mass & Hox (2005); Snijders (2005); Scherebaum & Ferreter (2009), Mathieu et al. (2012); Scherbaum & Pesner (2019)
- Dealing with Missing Data in Multilevel Research
  - Newman & Sin (2004); Grund et al. (2019)

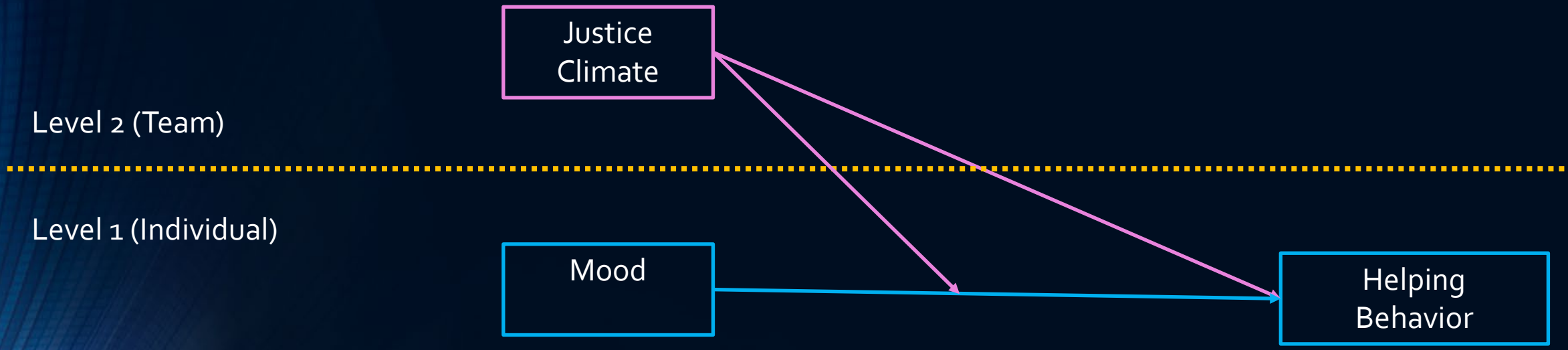
Advances in Multilevel Analyses

- Multilevel Regression: Raudenbush & Bryk (2002); Hox (2017)
- Dyadic Analyses – Atkins (2005); Kenny et al. (2006)
- Growth Models: Bliese & Ployhart (2002)
- Multilevel SEM: Preacher et al. (2010); Vandenberg & Richardson (2019)
- Analyses for Non-Normal Data – Zachary et al. (2019)
- Variance Partitioning – LaHuis et al. (2014); LaHuis et al. (2019)
- Computational Models – Newman & Wang (2019)

Growth in Reproducibility and Replicability



Multilevel Models

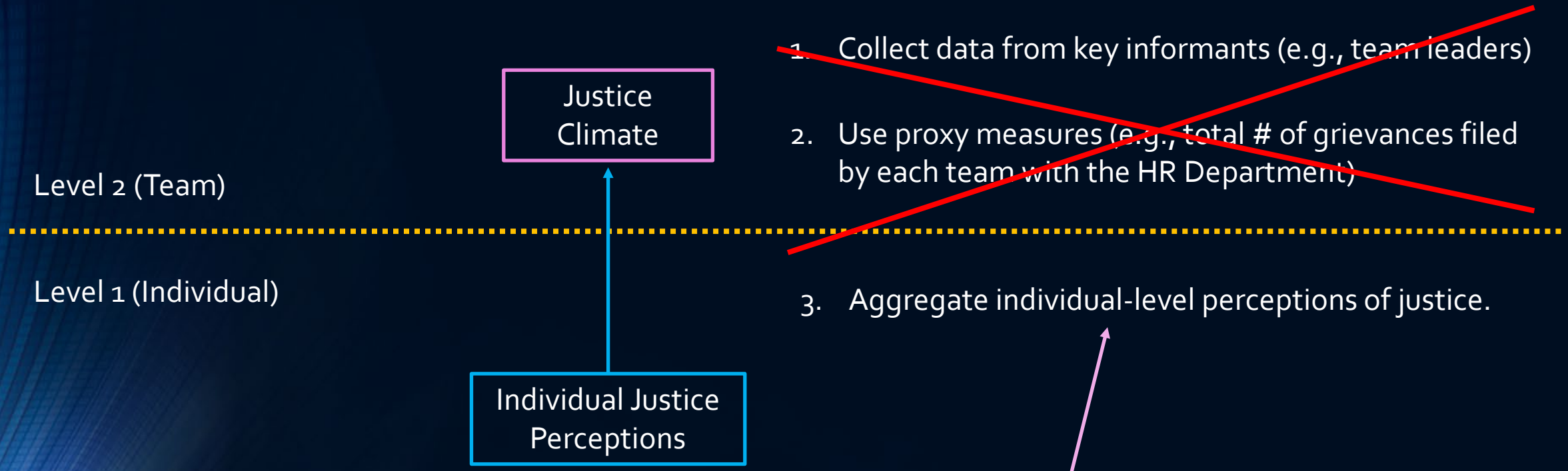


Hypothesis: Mood will be positively related to Helping Behavior.

Hypothesis: Justice Climate will be positively related to Helping Behavior.

Hypothesis: Justice Climate will moderate the strength of the relationship between Mood and Helping Behavior, such that, the positive relationship becomes stronger as scores on Justice Climate increase.

Measurement of Justice Climate



1. Collect data from key informants (e.g., team leaders)
2. Use proxy measures (e.g., total # of grievances filed by each team with the HR Department)
3. Aggregate individual-level perceptions of justice.

The construct of Team Climate is an upward, emergent phenomena.

Emergence

- A phenomenon is emergent when it originates in the cognition, affect, behaviors, or other characteristics of individuals, is amplified by their interactions, and manifests as a ***higher-level collective phenomenon*** (p. 55, Kozlowski & Klein, 2000; emphasis added).

Measurement Models for Emergent Phenomena

- Consensus Models
  - Convergent Emergence (Kozlowski & Klein, 2000)
  - Pooled Constrained Emergence (Kozlowski & Klein, 2000)
  - Direct Consensus (Chan, 1998)
- Non-Consensus Models
  - Pooled Unconstrained Emergence (Kozlowski & Klein, 2000)
  - Additive (Chan, 1998)
  - Dispersion (Chan, 1998)

Statistical Justification for Data Aggregation

- r_{WG} and $r_{WG(J)}$ – estimate of within-unit agreement/consensus; estimates range from 0 (complete lack of agreement) to 1 (perfect agreement)
- ICC(1) – estimate of non-independence; interpreted as the proportion of variability in scores obtained from lower-level units (e.g., individuals) that may be attributed to the nesting of those lower-level units within higher-level units (e.g., teams)
- ICC(2) – estimate of the the stability of unit-level means; group means computed on larger groups will be more reliable than means computed on smaller groups

Interrater Agreement: r_{WG}

- Formula:

$$r_{WG} = 1 - \frac{S_X^2}{\sigma_E^2}$$

- Where,

  S^2_X = observed variance among raters on variable X for a single target

  σ^2_E = expected variance on variable X when there is a complete lack of agreement -- akin to random responding

Interrater Agreement: r_{WG}

- Thus, S^2_X/σ^2_E represents the proportion of observed variance that is error variance engendered by random responding.
- Subtracting this ratio from 1 yields an estimate of the proportional reduction in error variance.
- When judges are in perfect agreement, $S^2_X = 0$ and $r_{WG(1)} = 1.0$.
- When judges lack perfect agreement, $S^2_X > 0$ and $r_{WG(1)} < 1.0$.

Interrater Agreement: $r_{WG(J)}$

- Formula:

$$r_{WG(J)} = \frac{J\left(1 - \frac{\overline{S_{Xj}^2}}{\sigma_E^2}\right)}{J\left(1 - \frac{\overline{S_{Xj}^2}}{\sigma_E^2}\right) + \left(\frac{\overline{S_{Xj}^2}}{\sigma_E^2}\right)}$$

- Where,

  $\overline{S_{Xj}^2}$ is the mean of the observed variances across *J* essentially parallel items.

  σ^2_E = same meaning as before

ICC(1)

Using variance components from a one-way random effects ANOVA:

$$ICC(1) = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$$

Where,

τ_{00} is the between-groups variance in individual-level scores

σ^2 is the (pooled) within-groups variance in individual-level scores

ICC(2)

Using the variance components from a one-way random effects ANOVA:

$$ICC(2) = \frac{\tau_{00}}{\tau_{00} + \sigma^2/k}$$

τ_{00} is the between-groups variance in individual-level scores

σ^2 is the (pooled) within-groups variance in individual-level scores

k is the number of lower-level units (e.g., individuals) nested in a particular higher-level unit (e.g., team)

Statistical Justification for Data Aggregation

- r_{WG} and $r_{WG(J)}$: provide separate estimates of within-group agreement for each team.
- ICC(1): provides an overall effect size reflecting the degree of non-independence or the proportion of variance in individual scores that may be attributed to their nesting in teams.
- ICC(2): provide separate estimates of the reliability for each group's mean; these statistics reflect the effectiveness of team means at distinguishing between the different teams.

Typical Description of Data Aggregation

A sample of 300 individuals nested in 50 work teams were asked to complete three measures. The 20 item Positive & Negative Affectivity Schedule (PANAS), the 10 item Organizational Citizenship Behavior (OCB) questionnaire, and the 10-item Workplace Justice Questionnaire (WJQ).

Useable data were available for 265 individuals nested in 42 teams.

We computed team-level Justice Climate scores as the mean of team members' scores on the WJQ. The decision to aggregate data was supported by estimates of within-group agreement (r_{WG} = 0.81; ICC=0.24) (Bliese, 2000; James et al., 1984; LeBreton & Senter, 2008).

- Is this clear?
- Is this transparent?
- Is it reproducible?
- What information is missing?

Missing Information

- How did the researchers conceptualize justice climate emergence? Why was their model an appropriate model?
  - Convergent emergence model, pooled constrained model, direct consensus model, referent shift consensus model, etc.?
- r_{WG} is used with single-item measures. However, the WJQ is a 10-item measure so $r_{WG(J)}$ should have be estimated.
  - Was this just sloppy referencing? Or, was the wrong statistic computed?
- r_{WG} is estimated for each of the 50 teams. Thus, what exactly is represented by $r_{WG} = 0.81$?
  - What does 0.81 reflect? Is it referring to a mean or median of the values estimated across teams?

Important Omitted Information

- What null distributions were used to estimate r_{WG} and how/why were these distributions selected?
  - Rectangular? Skewed? Triangular?
- What were the actual criteria used to justify data aggregation?
  - $r_{WG} > 0.50$, or > 0.70, or > 0.90?
- Did all teams have sufficient agreement to justify aggregation?
  - If a team lacked agreement, how were data from those teams treated?
- What exactly does the ICC value represent?
  - Is this an ICC(1) or an ICC(2)?

Recommendations for Conducting & Reporting Data Aggregation

- Developed a set of Recommendations for Reporting Data Aggregation in Multilevel Research
- Compared our set of Recommendations to Current Practices.
- Reviewed 91 empirical articles published between 2017-2021.
- Sampled 6 prominent journals in the organizational sciences (JAP, AMJ, JOM, PPsych, JBP, JOB)

Recommendations vs. Current Practices

- **Recommendation #2:** Researchers should clearly identify the multilevel measurement model used as the basis for data aggregation (pooled unconstrained, pooled constrained, additive, direct consensus, etc.).

  - *Only 28% of the articles included this information*

Recommendations vs. Current Practices

- **Recommendation #3:** Researchers should include estimates of ICC(1) for all lower-level variables that are being aggregated to higher-levels.

  - *95% of the articles included this information*

Recommendations vs. Current Practices

- **Recommendation #5:** Researchers using $r_{WG}/r_{WG(J)}$ should include information about a) which null distributions were used to compute agreement, b) why those distributions were appropriate, and c) how/where the point-estimates of σ_E^2 were obtained.

  - *Only 25% of the articles included this information*

Recommendations vs. Current Practices

- **Recommendation #6:** Researchers should report multiple estimates of within-group agreement. Specifically, researchers are encouraged to include at least one estimate of agreement scaled on the 0 to 1 metric (r_{WG}/$r_{WG(J)}$, a_{WG}, $a_{WG(J)}$) and at least one estimate of agreement scaled on the original metric of the items (e.g., AD_M, SD).

  - Only ~2% (2 out of 91) papers included AD_M
  - Only ~3% (3 out of 91) papers included SD
  - None of the papers included a_{WG}

Recommendations vs. Current Practices

- **Recommendation #7:** Estimates of ICC(2) should be reported when unit-level means are computed to serve as aggregate-level variables.

  - 91% included this information

Recommendations vs. Current Practices

- **Recommendation #9:** Researchers should articulate the specific criteria used to determine whether data should be aggregated.

  - *Only 27% of the articles included this information*

Recommendations vs. Current Practices

- **Recommendation #10:** Researchers should describe the pattern of agreement across their data. This might include: a) descriptive statistics for the estimates of agreement, and b) a histogram to aid in visualizing the distribution of estimates.
  - *Only 65% of articles reported the mean r_{WG} value*
  - *Only 23% of articles reported the median r_{WG} value*
  - *Only 6% of articles reported the range of r_{WG} values*
  - *None of the papers included a histogram of r_{WG} values*

Recommendations vs. Current Practices

- **Recommendation #11:** Researchers should clarify if anomalous findings were obtained, and if so, how they were handled.
- Such findings might include "out-of-range" $r_{WG}/r_{WG(J)}$ and mixed patterns of agreement across groups (i.e., some having strong agreement, others not).

  - Only 10% of the articles (9 of 91) mentioned out-of-range values
    - 3 of the 9 reported no out-of-range values
    - 6 of the 9 reported out-of-range values; only 2 reported how those cases were handled

Summary

- Sizeable gaps exist between recommended reporting standards and current reporting practices.
- One explanation for this gap may be the (limited) availability of software packages for estimating within-group agreement statistics.
  - Option #1: Manually compute statistics in SPSS, SAS, or Excel.
  - Option #2: Paul Bliese's *multilevel* package in R – very user-friendly, maybe a bit too user-friendly ☺

wga:Within-Group Agreement

- *wga* (within-group agreement): a new R package
  - Extends prior R packages in this area (i.e.,Paul Bliese's *multilevel*).
  - Our package is a work-in-progress, but here is a preview of what we have so far…

```
fit <- WGA(x=lq2002[,c("LEAD01")],
           grpid=lq2002$COMPID,
           scale=c(1,5),
           model="Consensus",
           reset=T)
```

Item(s) to be aggregated

Grouping variable

Min & max values of response scale

Name of multilevel measurement model

Should out of range values be reset to zero?

wga:Within-Group Agreement

```
$data.aggreation.model
[1] "Consensus"

$wga.descriptives
           vars  n  mean    sd median trimmed   mad   min   max range  skew kurtosis   se
grp.size      1 49 41.67 27.57  30.00   39.73 25.20 10.00 99.00 89.00  0.52    -1.17 3.94
num.items     2 49  1.00  0.00   1.00    1.00  0.00  1.00  1.00  0.00   NaN      NaN 0.00
item.var      3 49  1.15  0.30   1.16    1.16  0.25  0.41  1.76  1.35 -0.35    -0.15 0.04
rwg.un        4 49  0.42  0.15   0.42    0.42  0.12  0.12  0.80  0.68  0.37    -0.10 0.02
rwg.ss        5 49  0.17  0.18   0.13    0.15  0.19  0.00  0.69  0.69  1.10     0.45 0.03
rwg.ms        6 49  0.04  0.12   0.00    0.01  0.00  0.00  0.54  0.54  2.86     7.42 0.02
rwg.hs        7 49  0.00  0.01   0.00    0.00  0.00  0.00  0.07  0.07  6.58    42.12 0.00
rwg.tri       8 49  0.16  0.18   0.12    0.14  0.18  0.00  0.69  0.69  1.18     0.63 0.03
rwg.nor       9 49  0.07  0.15   0.00    0.04  0.00  0.00  0.61  0.61  2.27     4.27 0.02
awg          10 49  0.43  0.14   0.43    0.42  0.13  0.13  0.76  0.63  0.33    -0.20 0.02
AD.mean      11 49  0.86  0.16   0.88    0.87  0.13  0.30  1.10  0.80 -1.04     1.72 0.02
AD.median    12 49  0.81  0.18   0.82    0.82  0.15  0.23  1.07  0.84 -0.95     1.09 0.03
```

wga:Within-Group Agreement

```
$rwg.out.of.range
  num.oor.un num.oor.ss num.oor.ms num.oor.hs num.oor.tri num.oor.nor reset.to.zero
           0          0         12         41          48          13           Yes

$rwg.error.variances
  scale.points uni   ss  ms   hs  tri  nor
             5   2 1.34 0.9 0.44 1.32 1.04

$wga.percentiles
$wga.percentiles$rwg.un
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
0.120 0.236 0.306 0.360 0.380 0.420 0.440 0.486 0.520 0.618 0.800

$wga.percentiles$awg.un
   0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
0.130 0.248 0.336 0.350 0.372 0.430 0.450 0.490 0.544 0.612 0.760
```

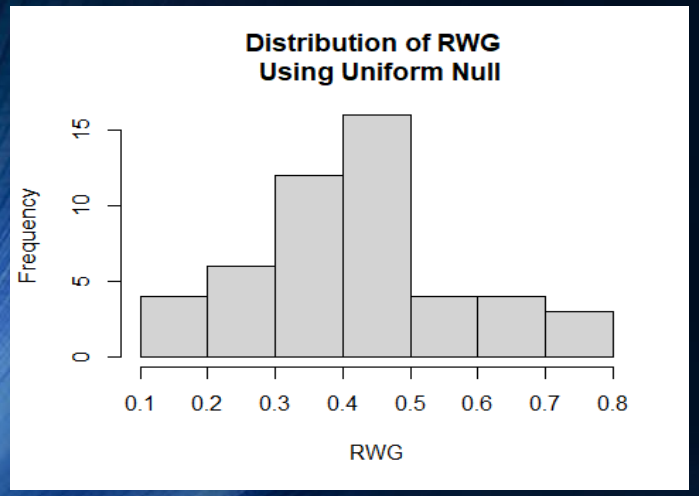
wga:Within-Group Agreement

$wga.results

	grp.name	grp.size	aggr.model	num.items	item.var	rwg.un	rwg.ss	rwg.ms	rwg.hs	rwg.tri	rwg.nor	awg
2	2	24	Convergent	1	1.26	0.37	0.06	0.00	0.00	0.05	0.00	0.40
3	3	37	Convergent	1	1.61	0.19	0.00	0.00	0.00	0.00	0.00	0.22
4	4	45	Convergent	1	1.12	0.44	0.16	0.00	0.00	0.15	0.00	0.45
5	5	58	Convergent	1	0.99	0.50	0.26	0.00	0.00	0.25	0.05	0.51
6	6	12	Convergent	1	0.70	0.65	0.48	0.22	0.00	0.47	0.33	0.61

…

$wga.plots



Better Description of Data Aggregation

A consensus emergence model was used to relate individual-level justice perceptions to team-level justice climate. This type of model requires establishing within-team agreement prior to aggregating scores (Chan, 1998; Kozlowski & Klein, 2000).

We estimated within-team agreement using $r_{WG(j)}$ based on a uniform null response distribution ($\sigma_E^2 = 2.00$; James et al., 1984). We selected this distribution because data were anonymous and thus unlikely to be influenced by systematic response biases. We set a minimum agreement threshold of $r_{WG(j)} > 0.51$, which reflects (at least) moderate within-team agreement (LeBreton & Senter, 2008; LeBreton et al., 2023).

Results supported decisions to aggregate data for most teams ($r_{WG(j)}$ Mean = 0.81, Median = 0.85, SD = 0.13). Specifically, we found that 40 of the 42 teams had $r_{WG(j)}$ values exceeding the minimum threshold, 35 teams had values exceeding 0.71, and 19 teams had values exceeding > 0.91. None of the observed $r_{WG(j)}$ estimates fell outside the normal range of 0 to 1.

Two teams had $r_{WG(J)}$ values falling below the minimum threshold. In order to provide the most stringent test of our hypotheses, these teams were removed from the data set. Within-team agreement was also estimated using the $AD_{(j)}$ (Burke et al., 1999) statistic, which yielded results that were consistent with $r_{WG(j)}$ (see online supplement for additional details).

Intraclass correlations also indicated that individual-level justice perceptions varied across teams (ICC(1) = 0.24) and that team-level estimates of justice climate were fairly reliable (ICC(2) = 0.65).

Conclusion

- We hope our recommendations facilitate greater clarity, transparency, and reproducibility for multilevel research.
- We hope our new R package, *WGA*, helps to bridge the gaps between the recommended reporting standards and contemporary reporting practices.
- We welcome your thoughts, comments, criticisms, etc.
- Thank you for your time and thoughtful consideration of our work!

Thank You

- Special Thanks to:
  - Larry James for introducing me to multilevel research
  - Paul Bliese, Gilad Chen, Stephen Humphrey, Dina Krasikova, John Mathieu, Bob Vandenberg, Fran Yammarino (and many others) for taking the time to chat with me about multilevel research
  - George Banks & Ernest O'Boyle for challenging me to think more critically about issues related to research transparency & reproducibility
  - Amanda Moeller & Jenell Wittmer for their many important contributions to this project
  - Steven Rogelberg and his editorial team for their timely and constructive feedback on earlier iterations of this project
  - Larry Williams for allowing us the opportunity to share our work with you

References & Resources

- LeBreton, J.M., Moeller, A.N. & Wittmer, J.L.S. Data Aggregation in Multilevel Research: Best Practice Recommendations and Tools for Moving Forward. J Bus Psychol 38, 239–258 (2023). https://doi-org.ezaccess.libraries.psu.edu/10.1007/s10869-022-09853-9
- https://github.com/james-lebreton/WGA