Wraparound Team Observation Measure: Inter-rater Reliability, Concurrent Validity, and Related Psychometrics

Ericka Wiggins, MA and Spencer Hensley
University of Washington School of Medicine, Seattle, WA

Jesse Suter, PhD
Center on Disability & Community Inclusion, University of Vermont

Eric Bruns, PhD
University of Washington School of Medicine, Seattle, WA

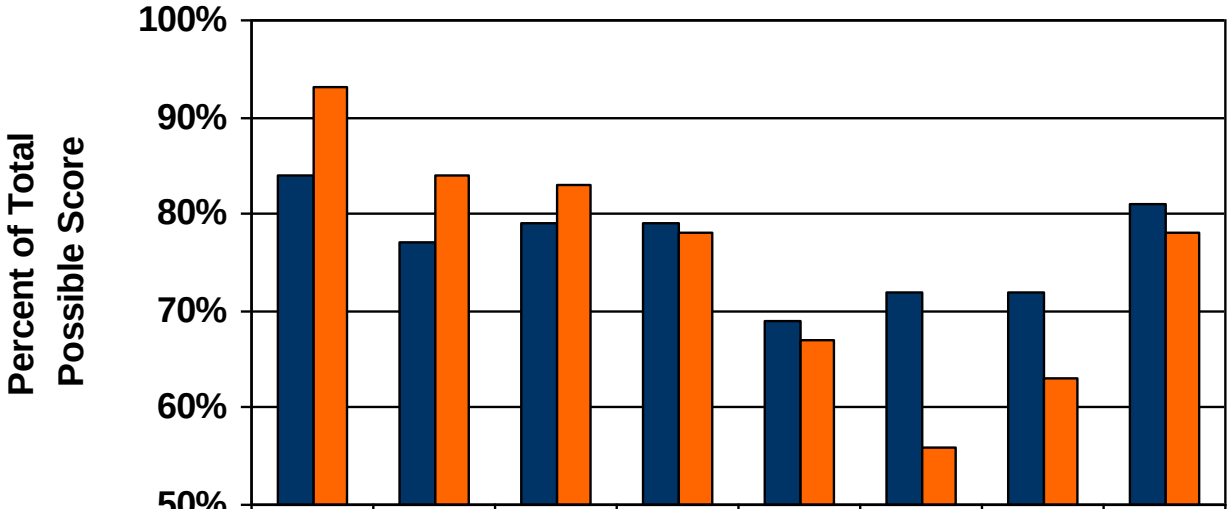
This project was supported by the Child, Adolescent, and Family Branch of the Center for Mental Health Services, SAMHSA
by the National Institute for Mental Health (R41MH077356; R34MH072759)
And by the Center on Disability & Community Inclusion at the University of Vermont

Team Observation Measure (TOM)

- The TOM was designed in 2006 to assess adherence to standards of high-quality wraparound during wraparound team meetings.
- It is organized according to the 10 principles of wraparound, with two items dedicated to each wraparound principal.
- Each of the 20 items has 3-5 indicators (71 total), which must be scored:
  - Yes (This was observed)
  - No (This was not observed)
  - N/A (This is not applicable)
- Use of the TOM is supported by a training toolkit that includes a self-test of knowledge of scoring rules and training to criteria using an online video
- The TOM is also supported by an online data entry, scoring, and reporting system (WONDERS; see www.wrapinfo.org)

Previous Research

- Initial version of the TOM (78 indicators):
  - Adequate internal consistency across 20 items overall
  - Adequate internal consistency for 17 of 20 items
  - Low to moderate inter-rater agreement
  - This information was used to complete a revision of the TOM in 2009
- Current version of the TOM
  - Strong agreement at a site level with scores from Wraparound Fidelity Index (WFI-4) interviews



	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8
Wrap Fidelity Index	84%	77%	79%	79%	69%	72%	72%	81%
Team Observation	93%	84%	83%	78%	67%	56%	63%	78%

Relationship between WFI-4 and TOM from 8 sites in California

$r(8) = .86, p<.01$

Goals for today's presentation

- Summarize new research on TOM psychometrics, reliability, and validity
  1. New analysis of 2 previous inter-rater reliability studies
  2. Psychometrics (descriptives and Cronbach's *alpha*) from national sample of N=955 team observations conducted since 2009
  3. Concurrent validity: Multi-level modeling of relationship between WFI-4 (interviews) and TOM (team observations) from national sample

1. Inter-rater Reliability

2009 Study (initial version of TOM) Sample and Method

Sample

Paired raters attended 15 wraparound team meetings for six different families and four different wraparound facilitators in California.

Method

Five graduate students in the social work program at California State University, Chico were trained on the administration of the TOM and observations were conducted in November and December 2008.

Results- Highest Agreement

Highest Agreement (100%)
1a: Parent/caregiver is a team member and present at the meeting.
2a: Team meeting attendees are oriented to the wraparound process and understand the purpose of the meeting.
3a: There is a clear agenda or outline to the meeting, which provides an understanding of the overall purpose of the meeting and the major sections of the meeting.
3c: The facilitator has prepared needed documents and materials prior to the meeting.

Results- Lowest Agreement

Less Than 60% Agreement
7a: Team provides multiple opportunities for natural supports to participate in significant areas of discussion.
11e: Talk is well distributed across team members and each team member makes an extended or important contribution.
12c: Team members do not hold “side-bar” conversations that do not include youth and family.
13a: The team has or sets goals with objective measurement strategies.
13b: The team assesses goals/strategies using measures of progress
14b: The team sets or reviews progress on measureable outcomes that are included in the plan.
18d: Serious challenges are discussed in terms of finding solutions, not termination of services or sanctions for the family.

Lowest Agreement (40%)
5c: The team uses a structured process or procedure to generate options or choices.
5d: The facilitator leads a robust brainstorming process to develop multiple options to meet priority needs.

How to Assess Inter-rater Reliability?

- Cohen's Kappa Coefficient
  - Measures the degree of agreement between two raters.
  - Corrects for the likelihood of agreement by chance

$$K = \frac{P_o - P_e}{1 - P_e}$$

2009 Inter-rater reliability study Results-Kappa

- 9 of 78 indicators had agreement < 60%
- Mean Kappa across indicators was .464

Discussion

- Results of the initial inter-rater reliability study indicate highly significant inter-rater agreement across the 78 indicators of the TOM (Kappa=.464;p<.000).
- However, Landis & Koch (1977) propose that Kappa values from 0.41-0.60 indicate only “moderate” agreement; Kappa values over 0.60 indicate “substantial” agreement.
- Results of this study were used to revise scoring rules and eliminate indicators that were difficult for observers to score reliably, allowing the TOM to be more reliable and valid in research.

TOM Revision

- Based on the results of the previous inter-rater reliability study, the TOM was revised in 2009.
- The current version of the TOM includes 71 indicators (compared to 78).

Sample and Method

Sample

Paired raters attended 12 wraparound team meetings for 12 unique families in Nevada.

Method

A research coordinator and wraparound program manager were trained on the administration of the TOM and observations were conducted between October 2009 and February 2010.

Results

- Only 2 of 71 indicators showed agreement of < 60%
- Mean Kappa across indicators was .835

Discussion

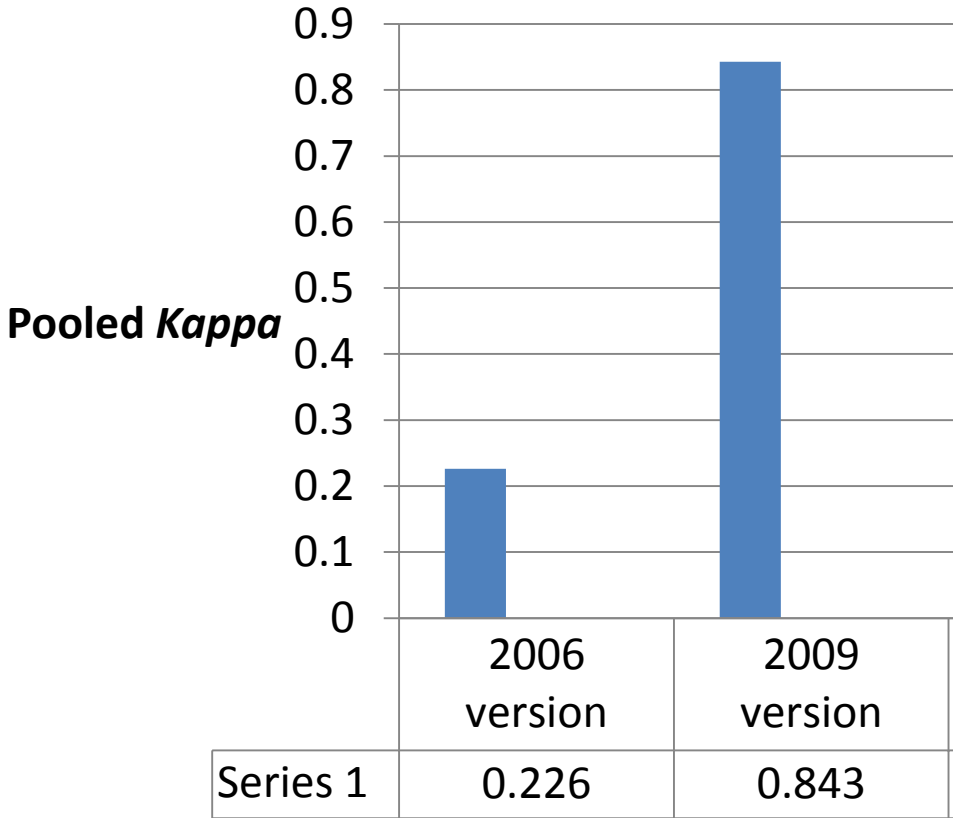
- This study suggested that inter-rater reliability of the TOM improved as a result of its revision.
- In a new analysis, we have calculated the pooled Kappa coefficient for the 2 versions of the TOM
  - Pooled Kappa is a more appropriate summary statistic for overall inter-rater reliability of a multi-item measure than merely averaging across all indicators

Pooled Kappa

- Pooled Kappa was calculated for the two previous studies.
- Pooled Kappa is the average of the averaged probabilities.

$$K_{pooled} = \frac{\overline{P}_o - \overline{P}_e}{1 - \overline{P}_e}$$

Results


Pooled *Kappa*
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2006 version
2009 version
Series 1
0.226
0.843

Conclusions

- Revision to the current version of the TOM improved inter-rater reliability
- Inter-rater reliability as found in this small scale study is “substantial” overall
- Currently conducting a new IRR study in a second site to add to the sample

2. Initial Psychometrics

Method

- Data were collected by local evaluators or supervisors trained to criteria using the *TOM Training Toolkit*
- July 2009 to August 2011
- Uploaded into WONDERS and compiled in de-identified fashion by the research team

Participants

10 Projects

- State (e.g., MA, PA)
- Large regions (e.g., LA County)

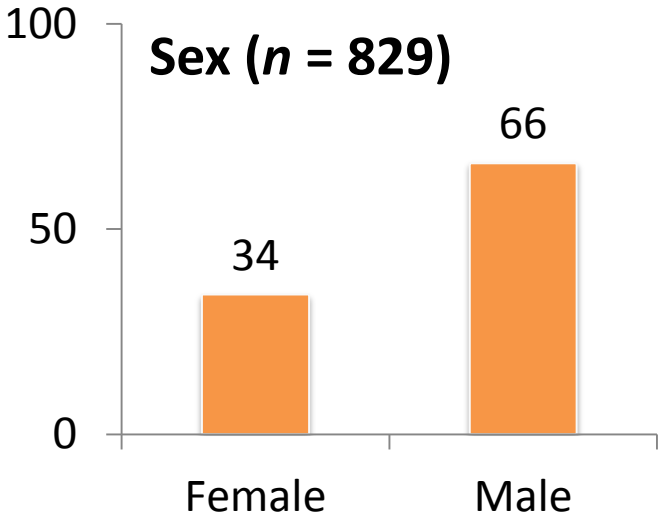
45 Sites

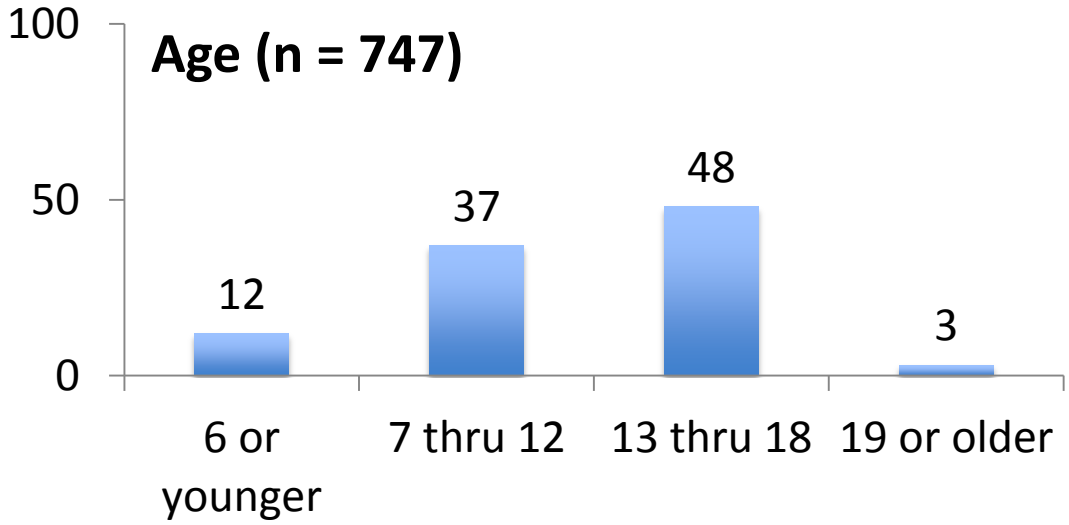
- Wraparound initiatives

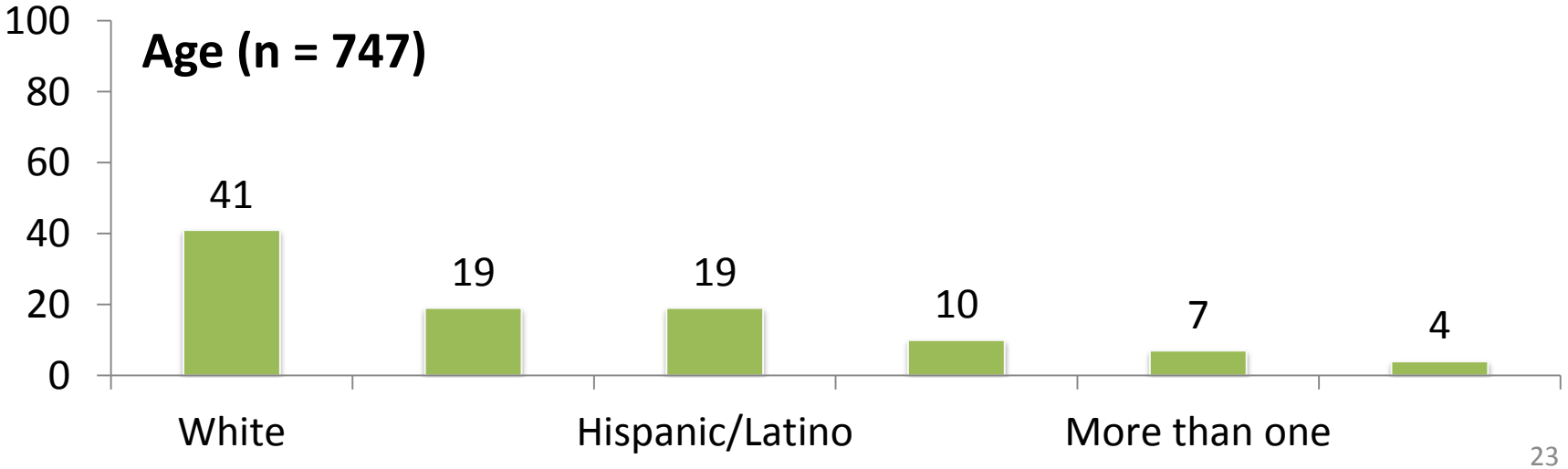
955 Teams

- Initial meetings (23%)
- Follow-up (73%)
- Transition & “Other” (4%)

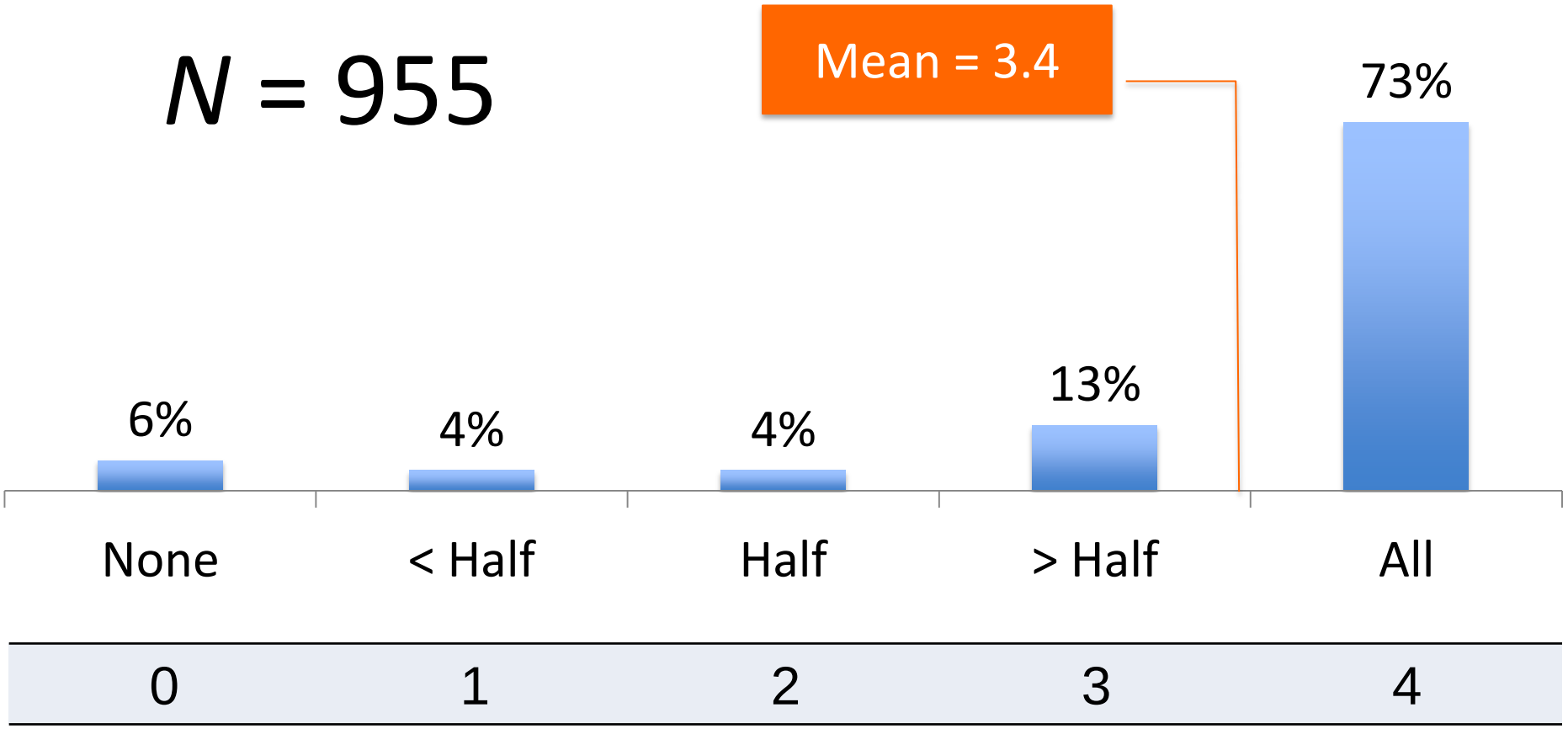
Youth Demographics (%)







TOM Number of Indicators Present



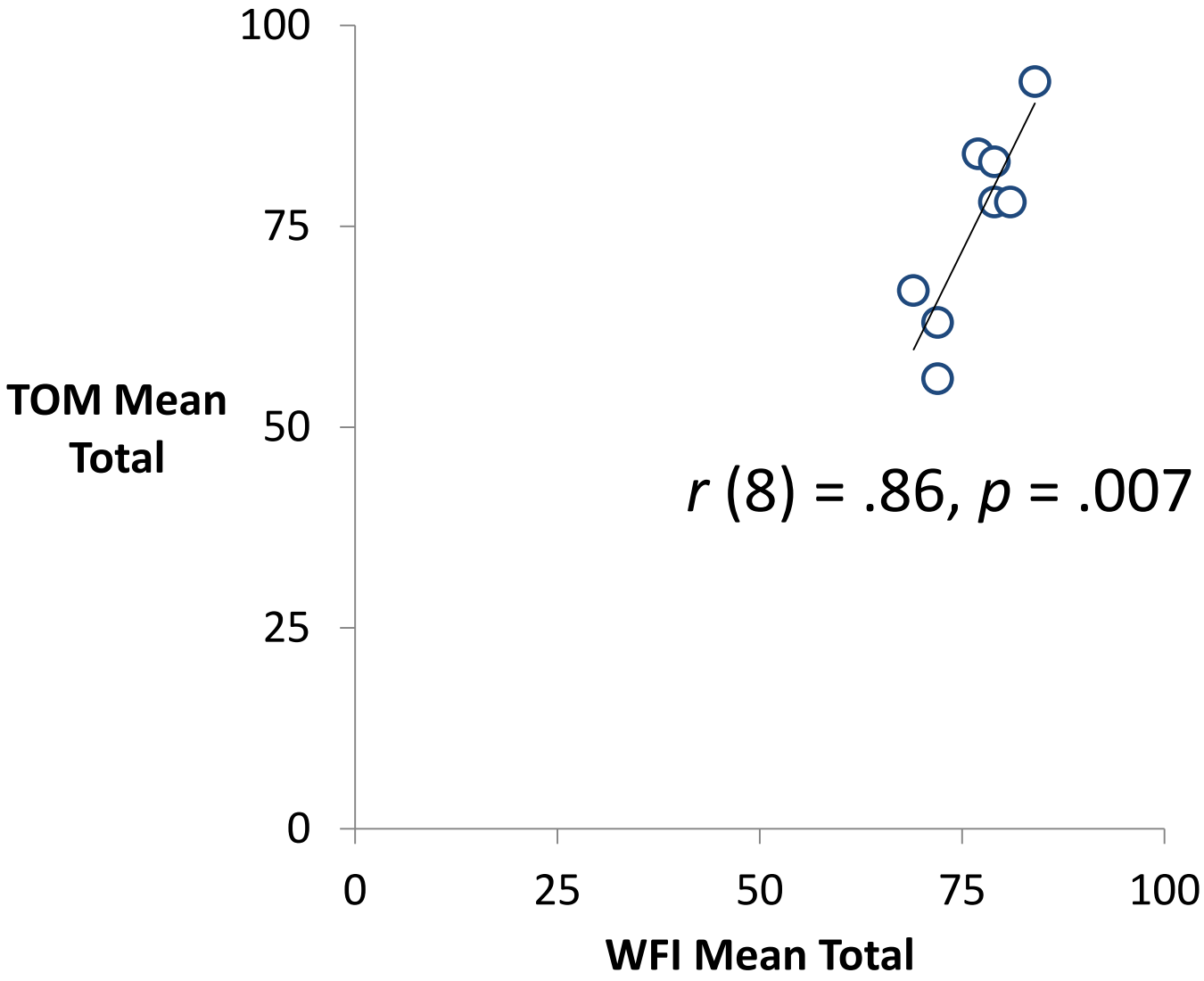
Total TOM Rating	M = 3.4	SD = 0.5

3. Concurrent Validity

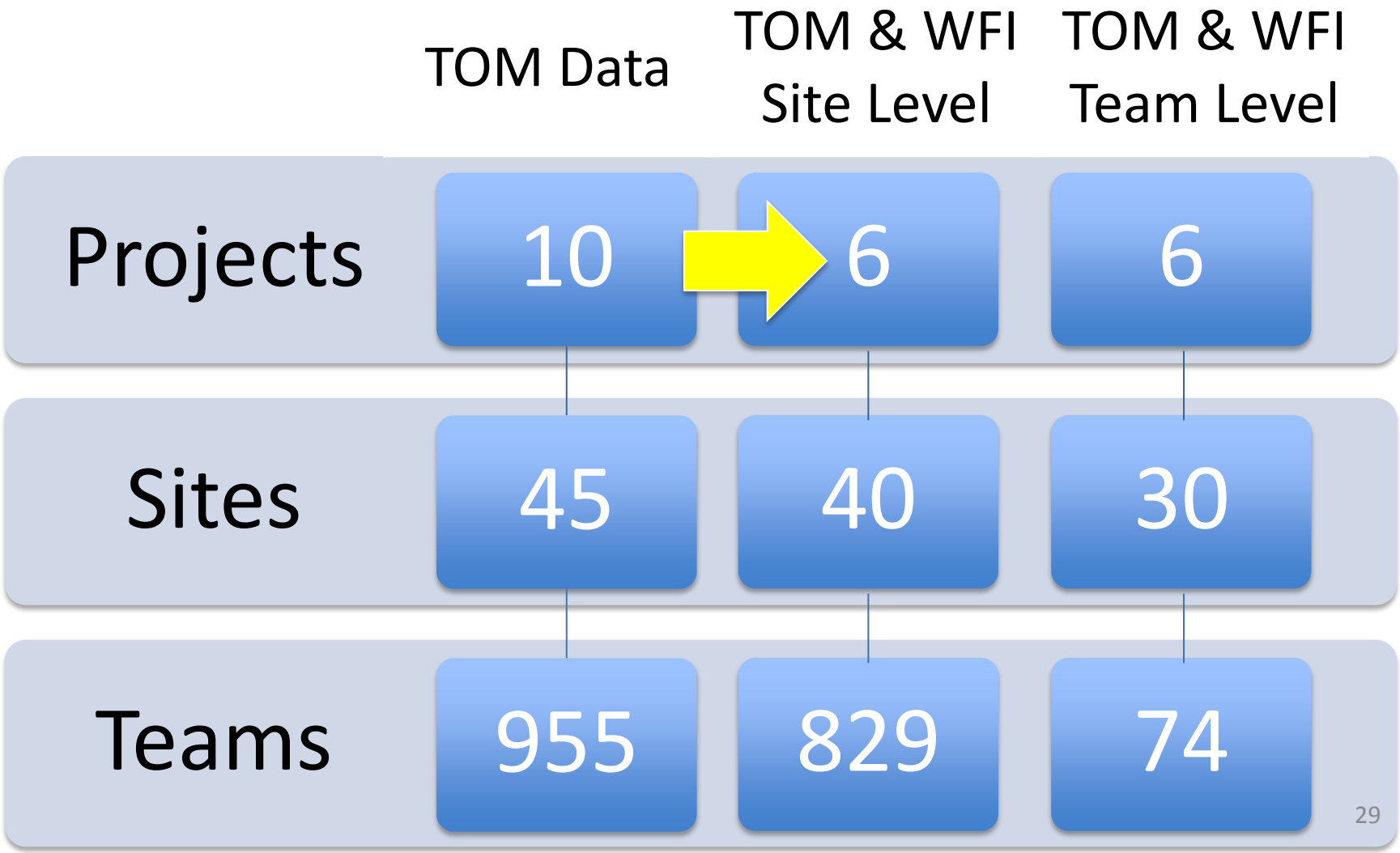
Concurrent Validity Study

- Comparison between TOM and WFI fidelity
- Validation of TOM and provide guidance to users of both measures
- Previous study showed strong relationship for using earlier version of TOM
- Current study uses revised version of TOM with comparisons at multiple levels

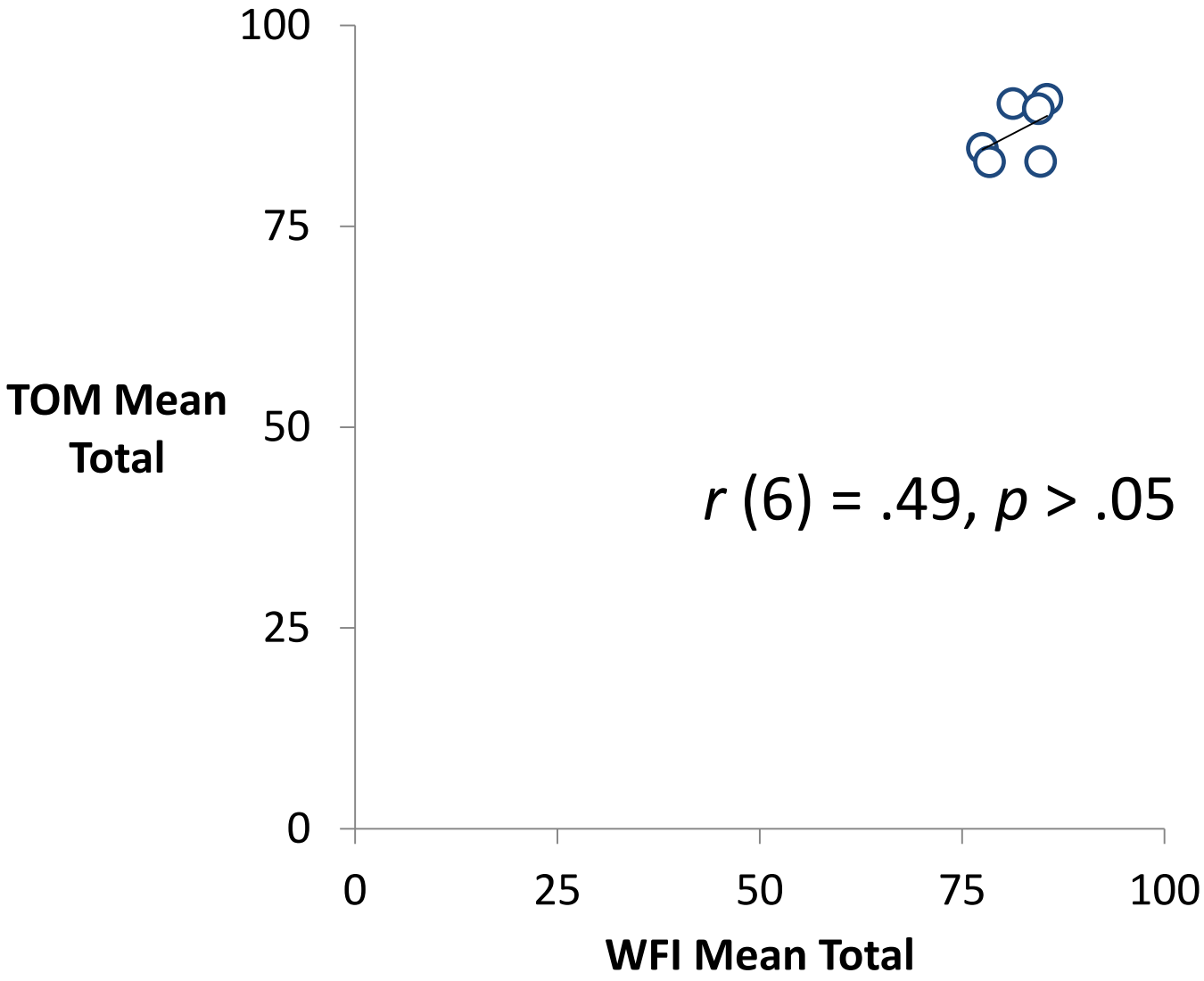
Previous Study



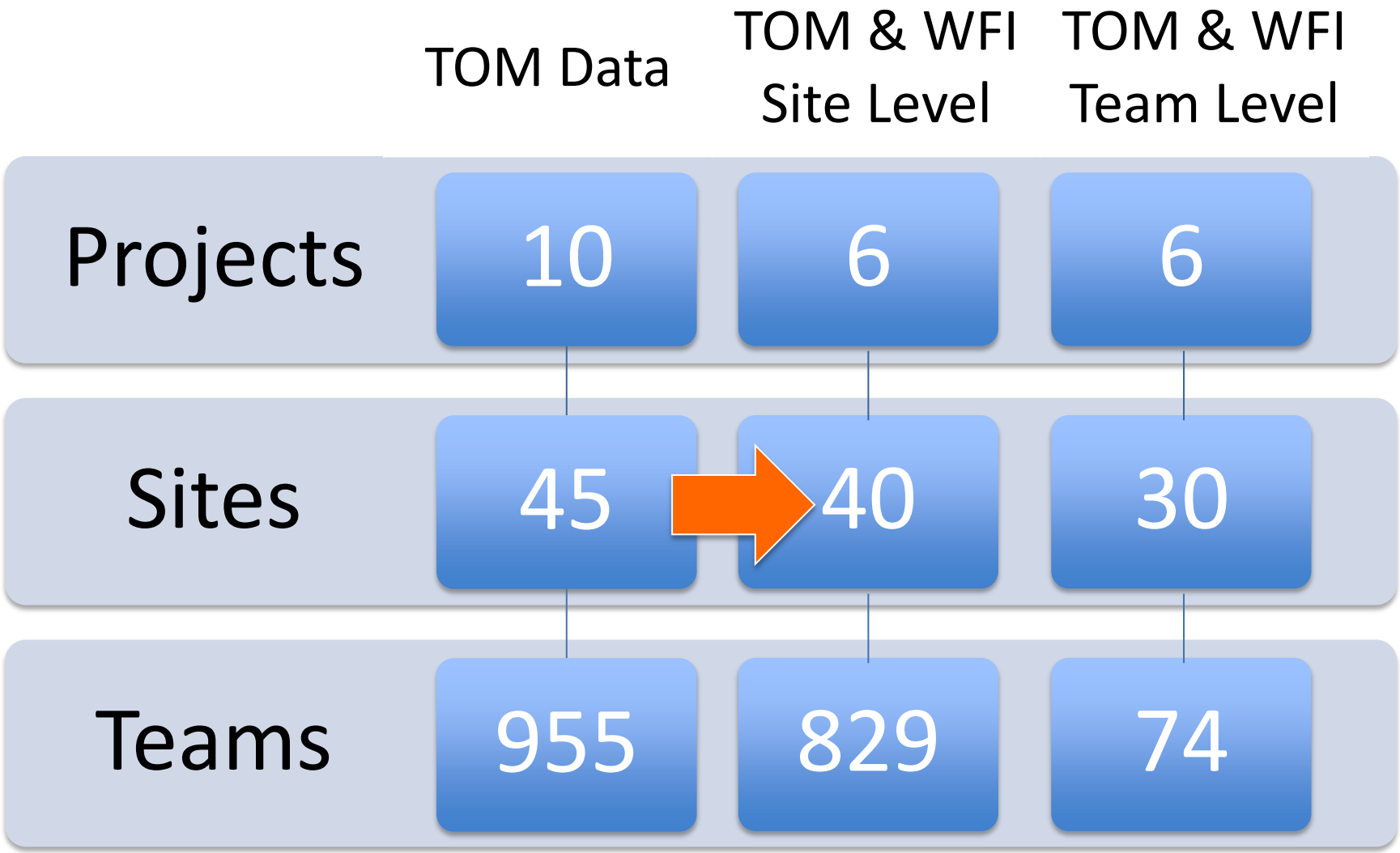
Nested Sample



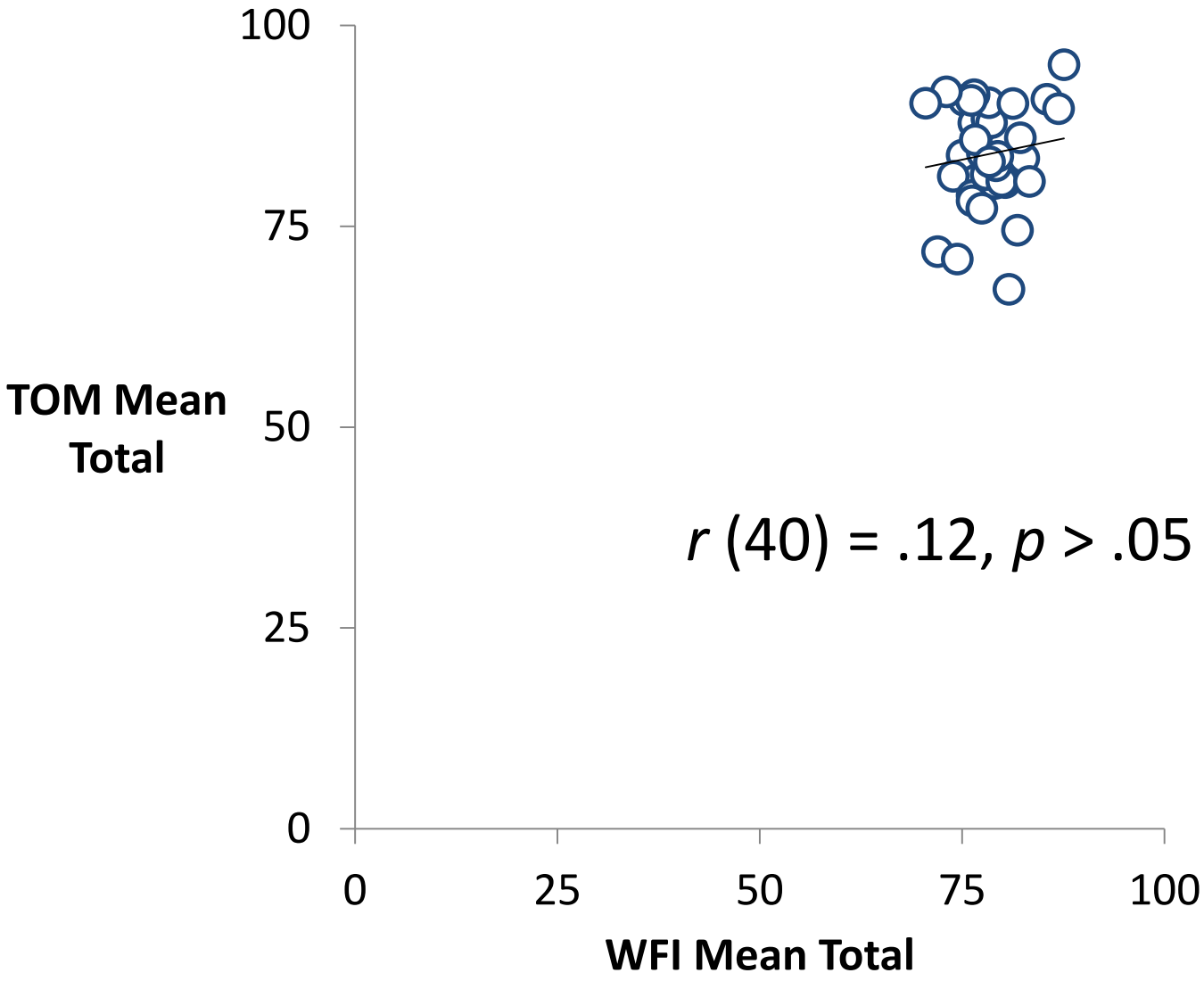
Project Level Correlation



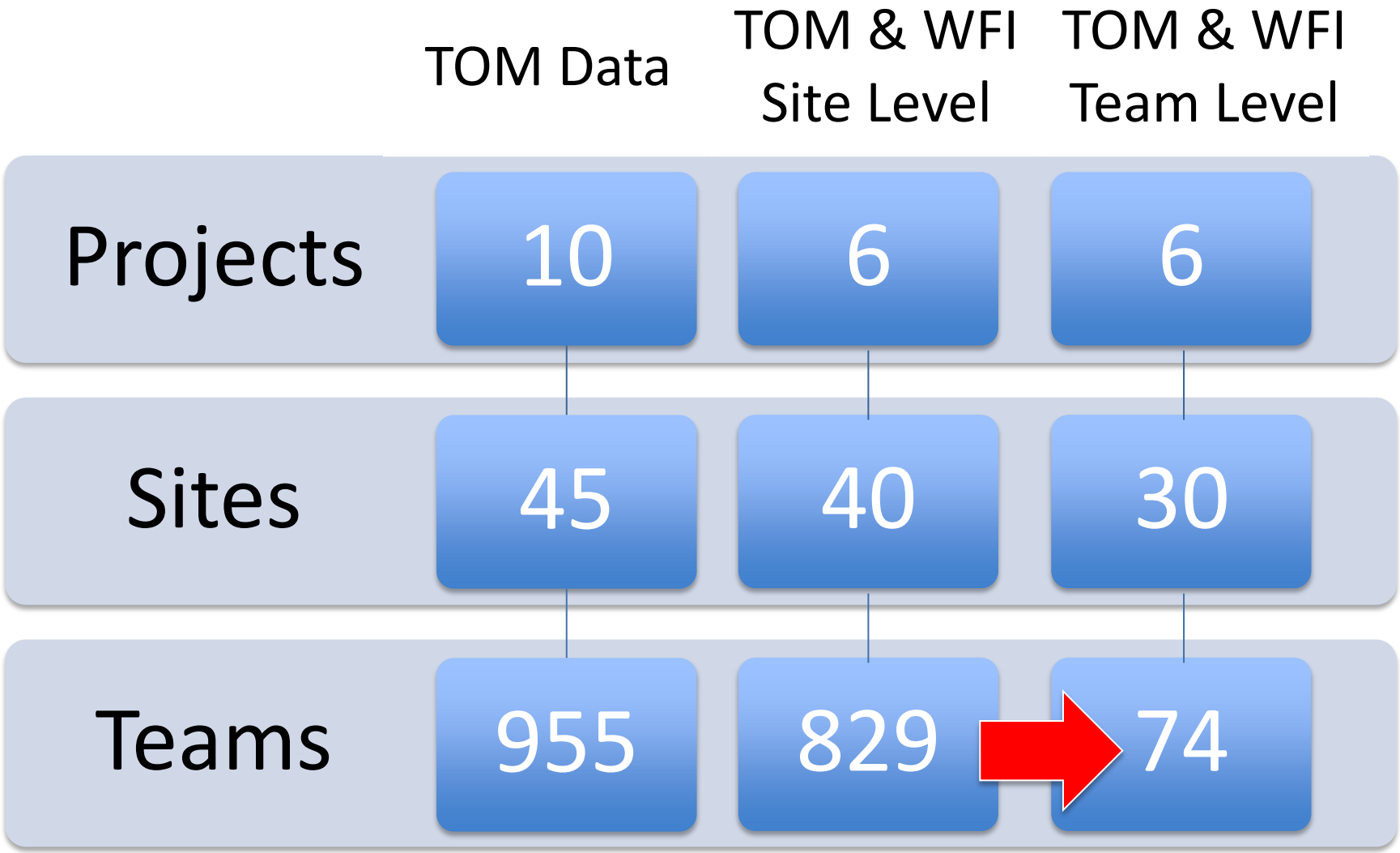
Nested Sample



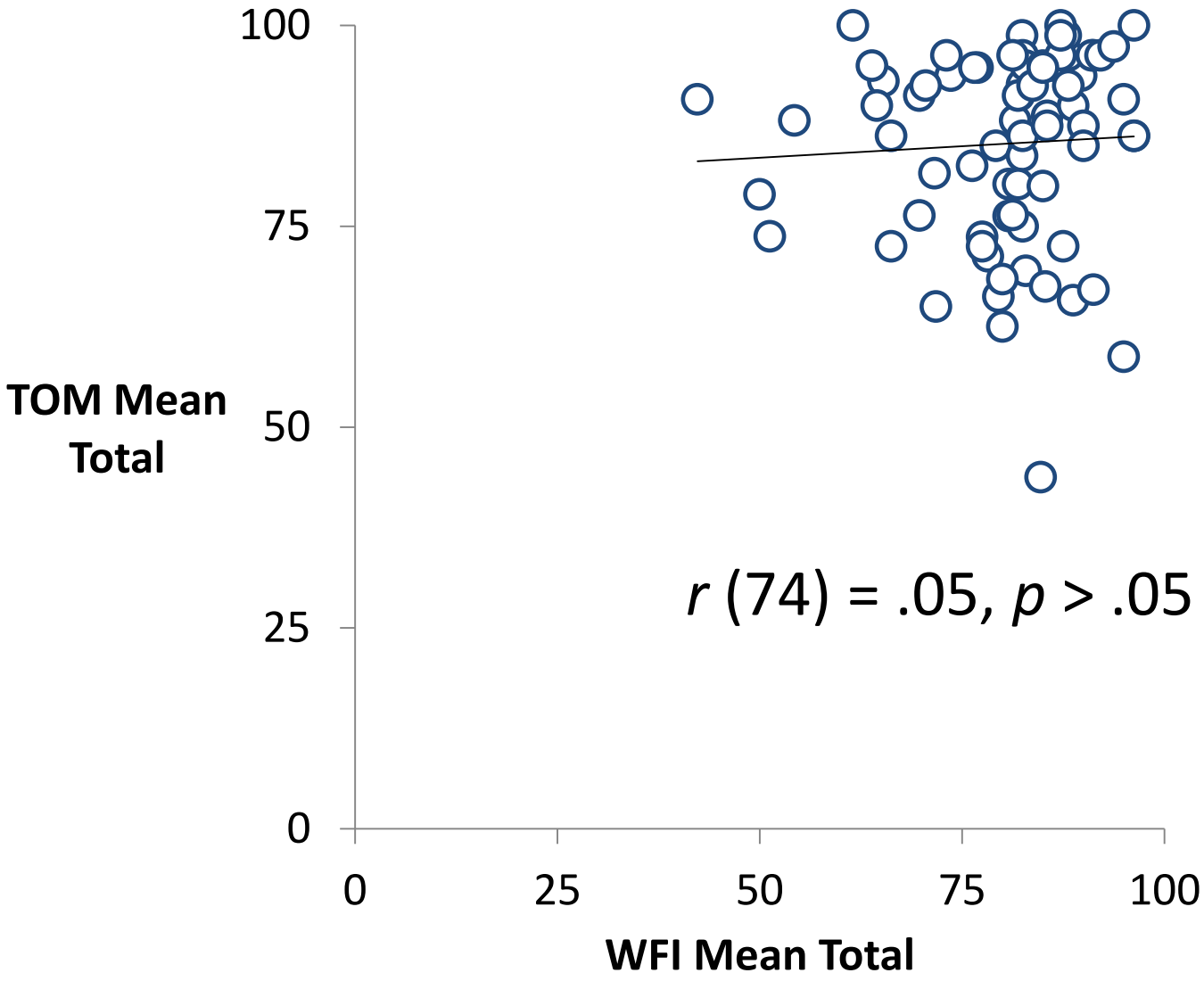
Site Level Correlation



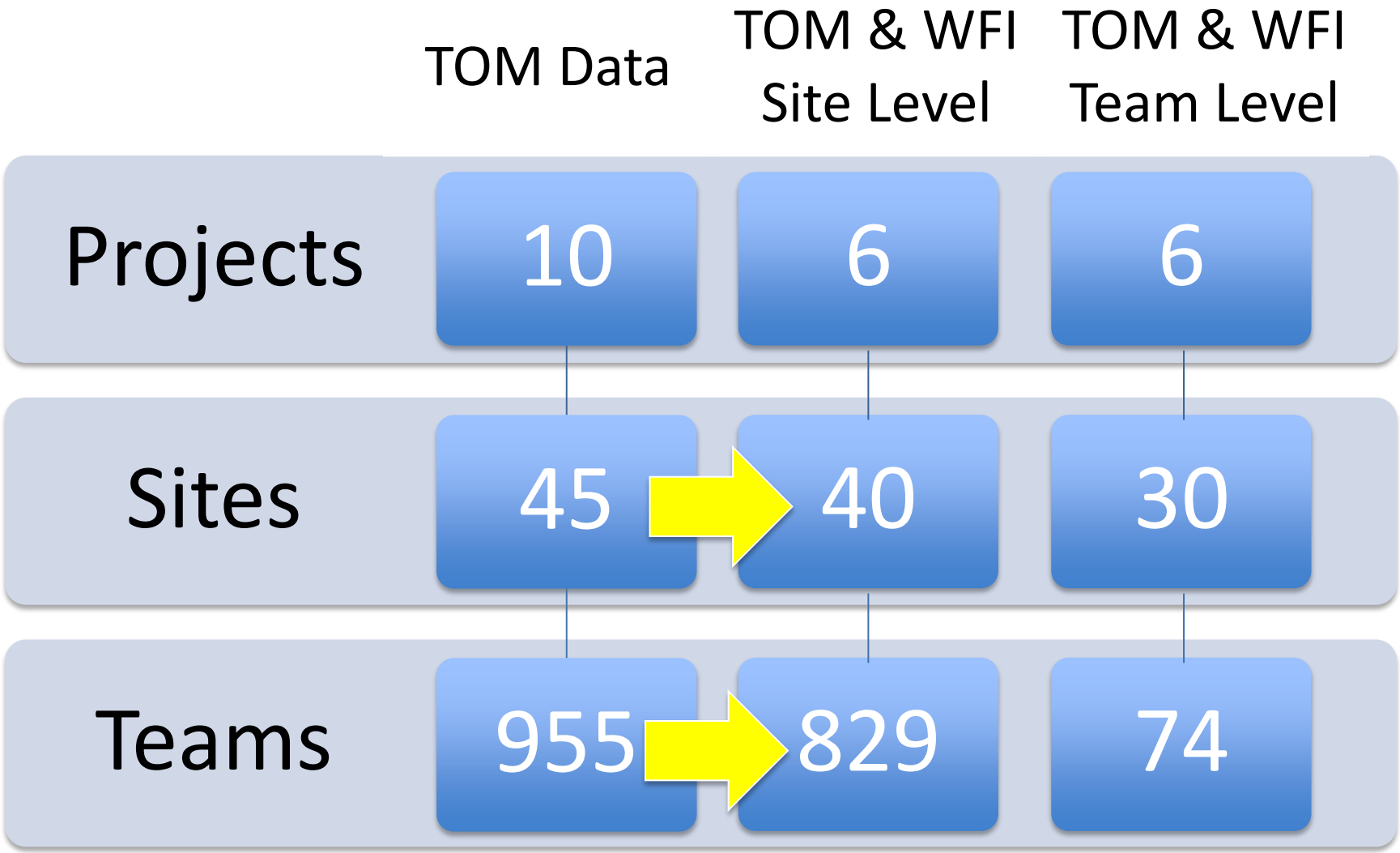
Current Study - Nested Sample



Team Level Correlation



Current Study - Nested Sample



Multi-Level Modeling

Parameter	Model 1	Model 2	Model 3
Intercept		-1.36 (.95)	-0.71 (1.06)
Level 1 (Team)			
Follow up meeting			.83 (.96)
Level 2 (Site)			
WFI Rating			.01 (.29)
External Observer			10.93 (7.1)
Random Effects			
Residual		*	111.78 (5.73)*
Intercept		5)	28.35 (8.48)*
Model Statistics			
AIC	7342.19	6322.82	6132.56
BIC	7351.91	6332.26	6141.93

No relationship found between TOM (team level) & WFI (site level) using HLM

Conclusions

1. Very high endorsement of wraparound indicators
   - TOM and WFI scores are increasing nationally every year
   - Reduces utility as a research tool
2. Stronger relationships at higher levels
   - TOM and WFI function as a valid agency or program level measures of fidelity
   - TOM and WFI may tap into very different things at an individual team/family level

Next Steps

Can TOM be improved through item analysis?

What factors at project & site level predict ratings?

Re-examine findings with controlled sample.