Studying NIH Grant Peer Review

Molly Carnes, MD, MS

Elizabeth Pier, PhD

September 1, 2017

NIH Director's Transformative R01
Exploring the Science of Scientific Review (R01 GM111002)

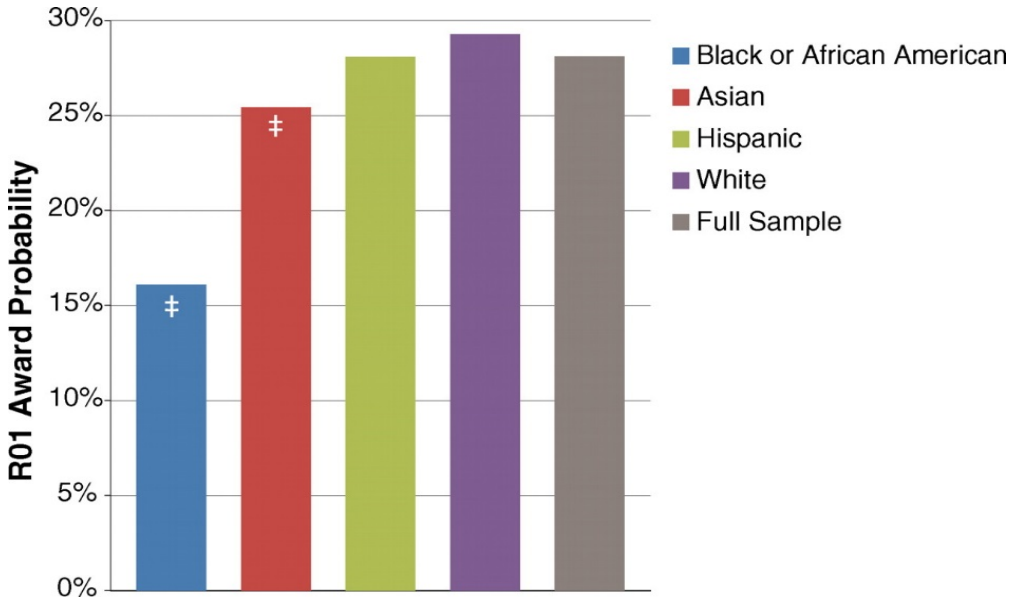
Multiple PI: M. Carnes, C. Ford, P. Devine

3 Studies:

1. Text analysis of R01 critiques.
2. Examination of constructed study sections.
3. Experimental manipulation of R01 applicant race and gender.

Race and Gender Differences in R01 Award Rates

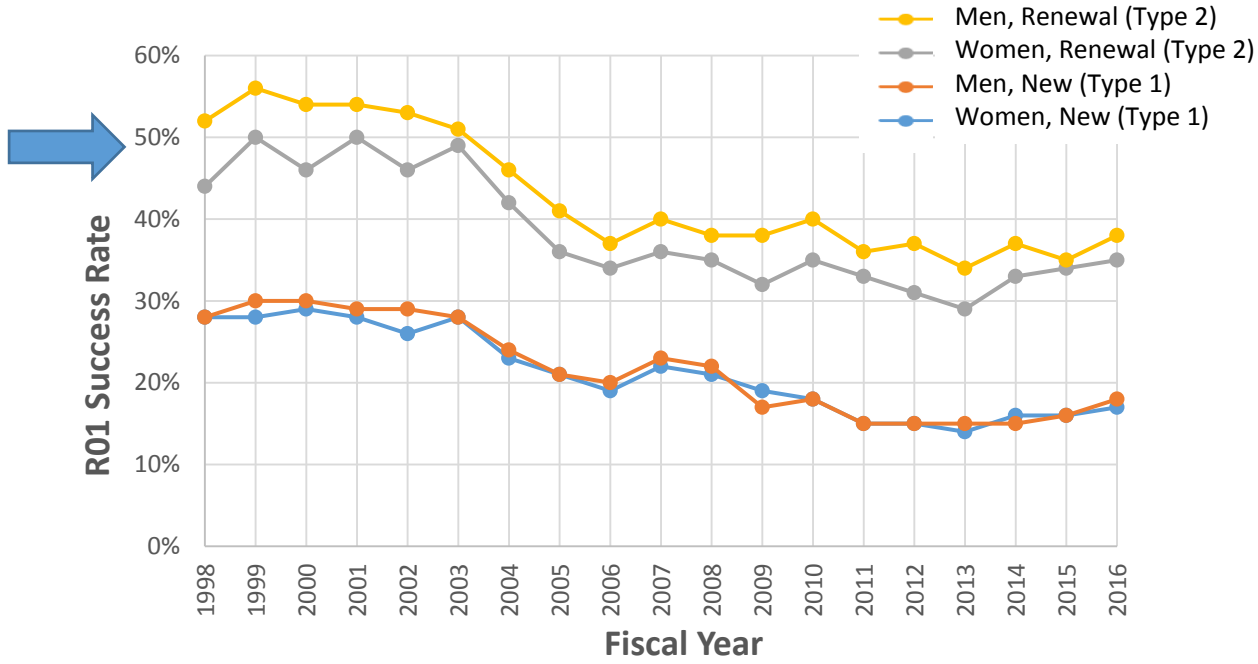
Black PIs have Lower R01 award probabilities than White PIs



Probability of NIH R01 award by race and ethnicity, FY 2000 to FY 2006 (N = 83,188); ‡, $p < .001$

Based on data from NIH IMPAC II, DRF, and AAMC Faculty Roster.
(Ginther, *Science,* 2011)

Female PIs have lower R01 renewal award rates than male PIs



Success Rates by Gender and Type of Application, FY 1998 to FY 2016

(NIH, 2017)

Subjectivity in evaluative criteria allows implicit assumptions to affect interpretation of objective data

- The temperature is 41 degrees F = objective
- Is that a cold or hot day? = subjective
  - Wisconsin = warm
  - Florida = cold

- R01 application submitted by a certain PI from a certain institution = objective
  - Innovative scientific leader
  - Pioneering research
  - Impact scores – 2, 3, 4?

} subjective

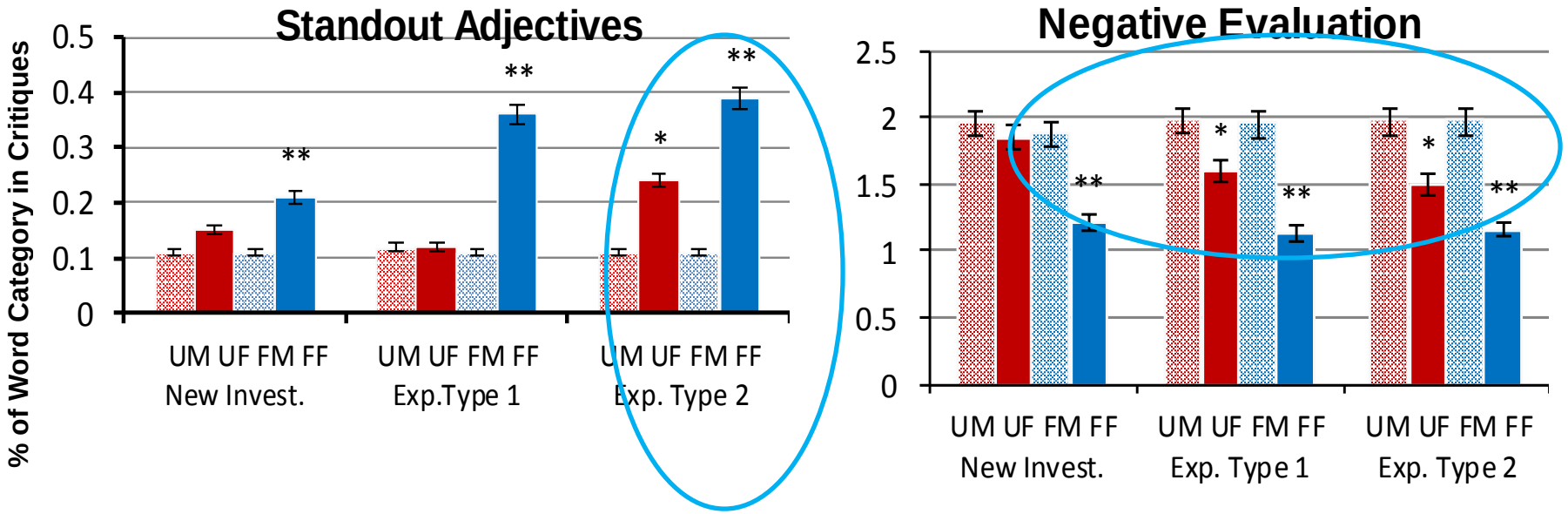
Stereotypes are known even if we don't believe them

Men[1]	Women[1]	White[2]	Asian[2]	Black[2]	Latino[2]
• Strong • Decisive • Stubborn • Competitive • Ambitious • Risk-taking • Assertive • Tough • Authoritative • Independent	• Caring • Nurturing • Bad drivers • Family-oriented • Communal • Supportive • Sympathetic • Nice • Helpful	• High status • Rich • Intelligent • Arrogant • Privileged • Blonde • Racist • All-American • Ignorant	• Intelligent • Bad drivers • Good at math • Nerdy • Shy • Skinny • Educated • Quiet	• Ghetto or unrefined • Criminal • Athletic • Loud • Gangsters • Poor • Have an attitude • Unintelligent • Uneducated	• Poor • Illegal immigrant • Uneducated • Family-oriented • Lazy • Day laborer • Unintelligent • Loud • Gangsters

1. Eagly and Sczesny (2009), Bem (1974); 2. Ghavami and Peplau (2015)

Quantitative text analysis of R01 critiques

- 443 grant reviews from R01s awarded after unfunded in 2008 (N=65)
- Women's: more standout adjectives (e.g., excellent, outstanding) ($p \leq 0.01$)
- Men's: more negative descriptors (e.g., unfocused, illogical) ($p \leq 0.01$)

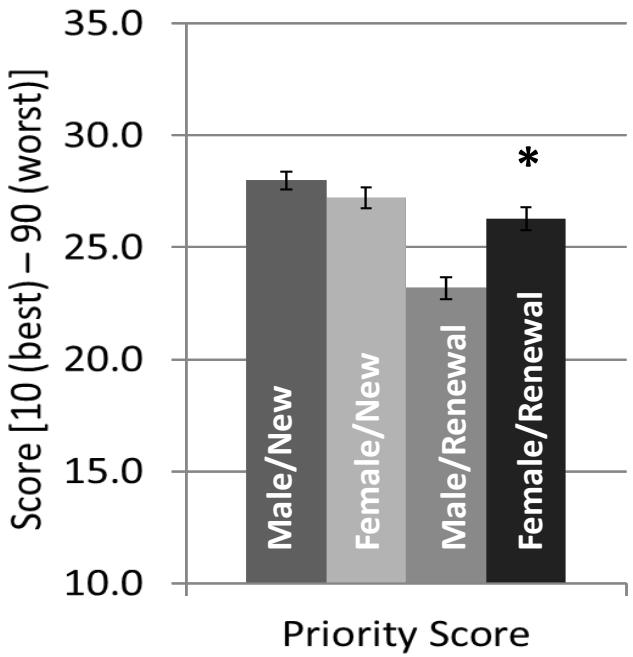


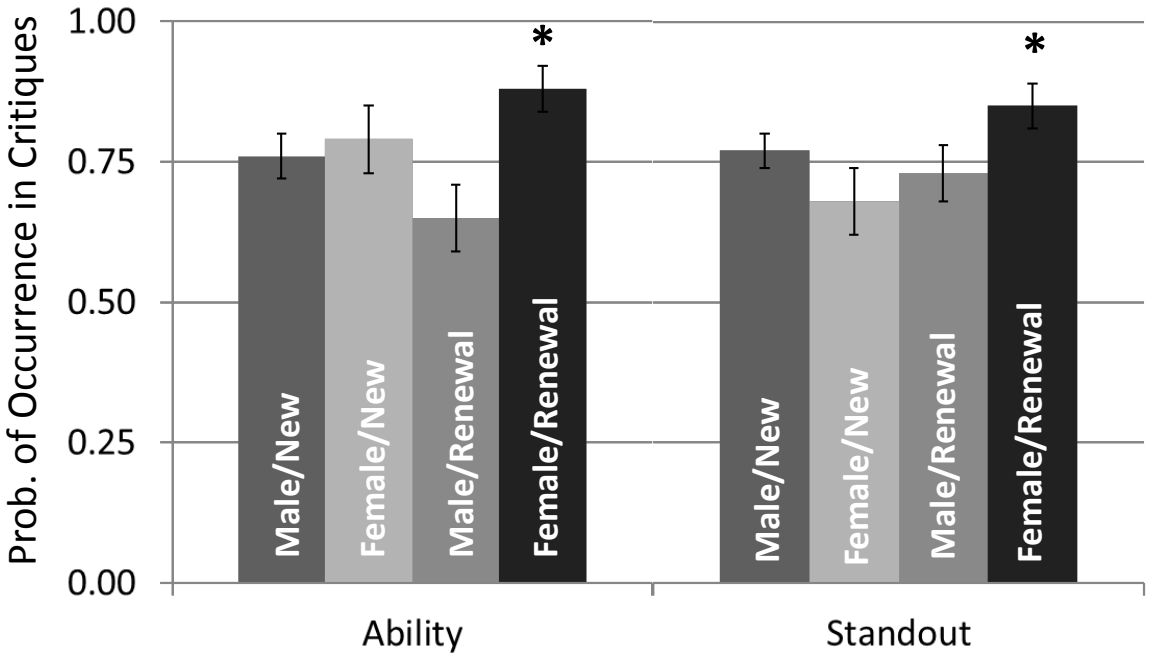
Do implicit assumptions lead to different referent standards in evaluating the applications of male and female PIs?

Kaatz et al., 2015

R01 Grant Critiques UW-Madison (N=739), 2010-2014: Greater praise for women ≠ better score

- Female PIs' R01 renewals assigned worse (higher) priority scores.
- More critiques of female PIs' R01 renewals included words about ability and standout adjectives (e.g., "outstanding," "excellent").





*Difference between groups is significant (P<.05); PIs in sample had similar levels of productivity and background qualifications; models controlled for funding outcome and experience level.

Kaatz et al., 2016

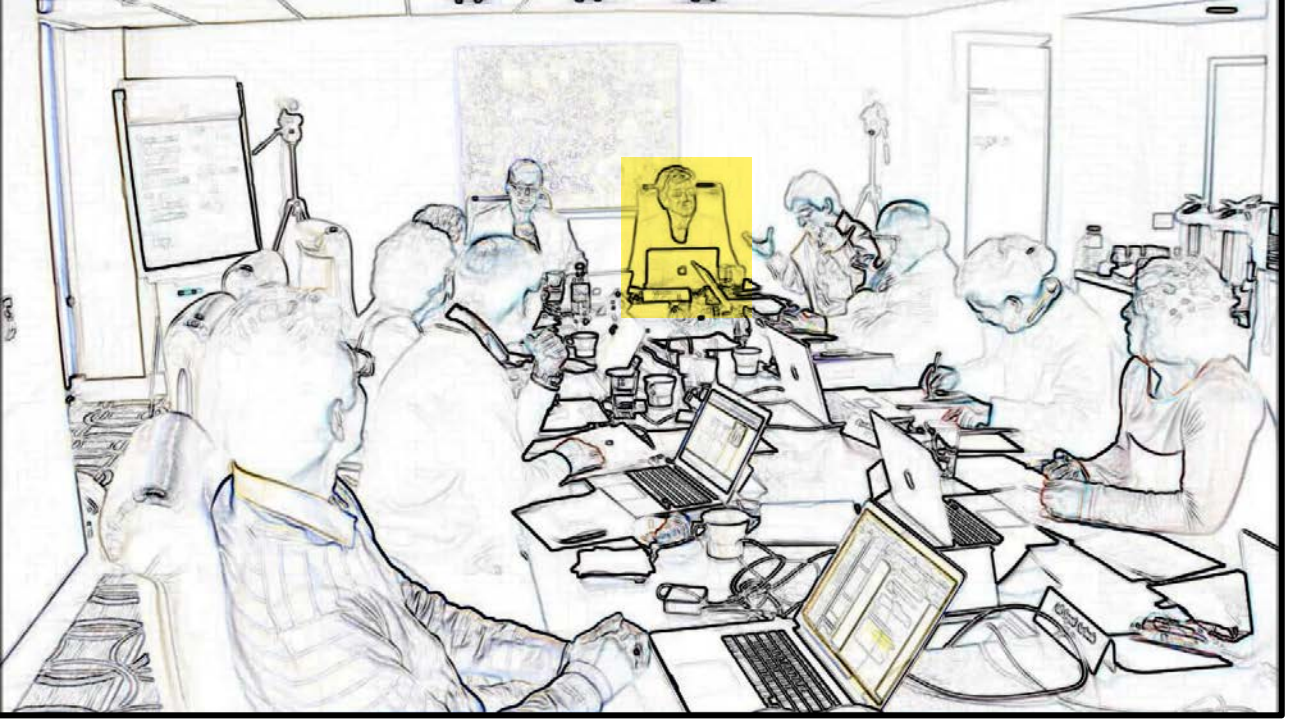
Summary of text analysis to date

- Our work suggests that characteristics of the investigator could introduce bias into evaluation of an application in ways that could contribute to the gender difference in Type 2 renewals
- Have collected a national sample of over 6000 PIs (~19,000 critiques) that will enable to include race in our analyses
- Examining algorithmic text mining approaches

Examining Constructed Study Section Meetings

- Subjectivity in evaluative criteria allows implicit assumptions to affect interpretation of objective data

Overall Impact	*Score*	*Descriptor*
High	1	Exceptional
	2	Outstanding
	3	Excellent
Medium	4	Very Good
	5	Good
	6	Satisfactory
Low	7	Fair
	8	Marginal
	9	Poor

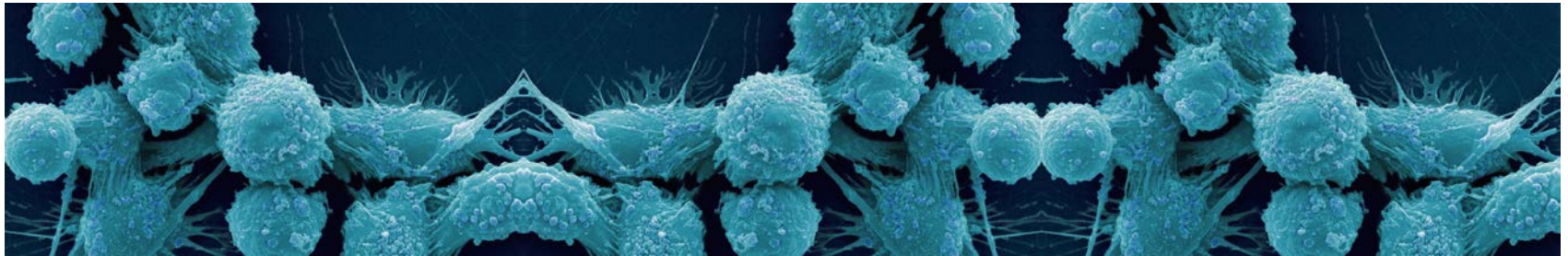


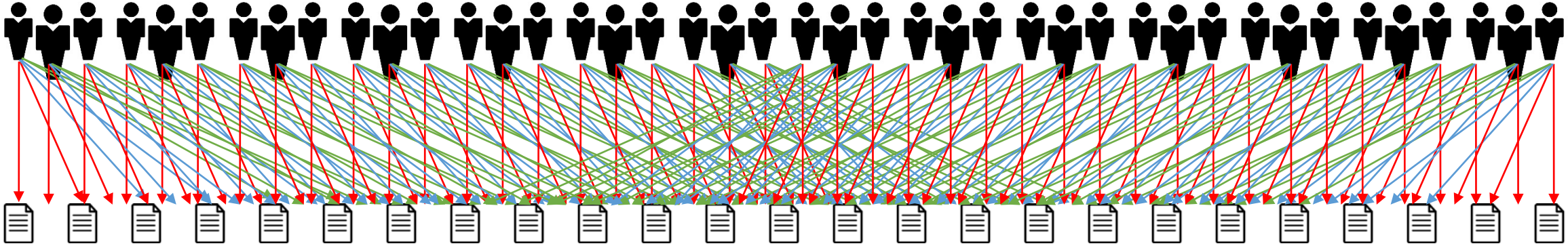
42

2012

2015

25






CSS 1


CSS 2


CSS 3


CSS 4

Application	CSS1	CSS2	CSS3	CSS4
Wu		☒	☒	●
Lopez	●	●	●	
Holzmann	☒		☒	●
Zhang	☒	☒	●	☒
Adamsson	●	☒	☒	☒
McMillan		☒	●	
Phillips	●	●	☒	
Amsel	●	●	●	
Abel	●	●	●	☒
Ferrera	☒	☒	●	☒
Stavros	☒	●	☒	●
Washington	●	●	☒	●
Williams	●	☒	●	●
Rice	☒	●	●	☒
Albert	●	☒	☒	●
Edwards	☒	●	☒	☒
Foster	●	●	●	●
Henry	●	●	●	●
Bretz	☒	☒	●	
Molloy	●	●	☒	
Wei	☒	☒	☒	☒
McGuire		☒	☒	
Kim	☒	☒	☒	
Bernard	☒	☒		
Lukska	☒	☒	☒	☒

● = Discussed
☒ = Triaged Out
Blank = Not Assigned

Scoring Variability

How variable are reviewers' scores?

Krippendorff's α

(Hayes & Krippendorff, 2007)

α ≥ 0.8	"Reliable"
α ≥ 0.67	"Tentative but not definitive"

Low agreement among individual reviewers

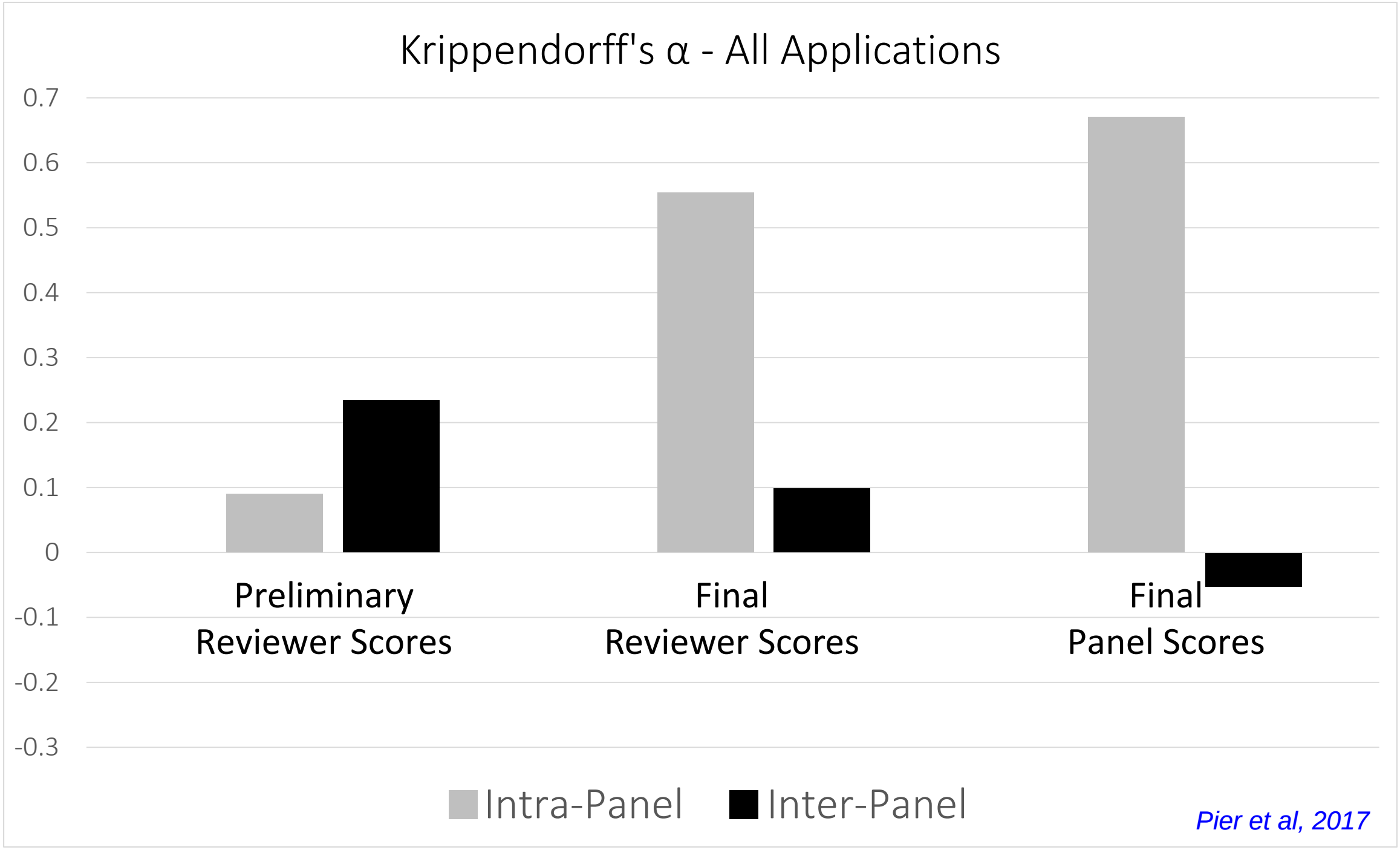
α = .0840

Better agreement ***within*** panels after collaborative discussion

α = .6652

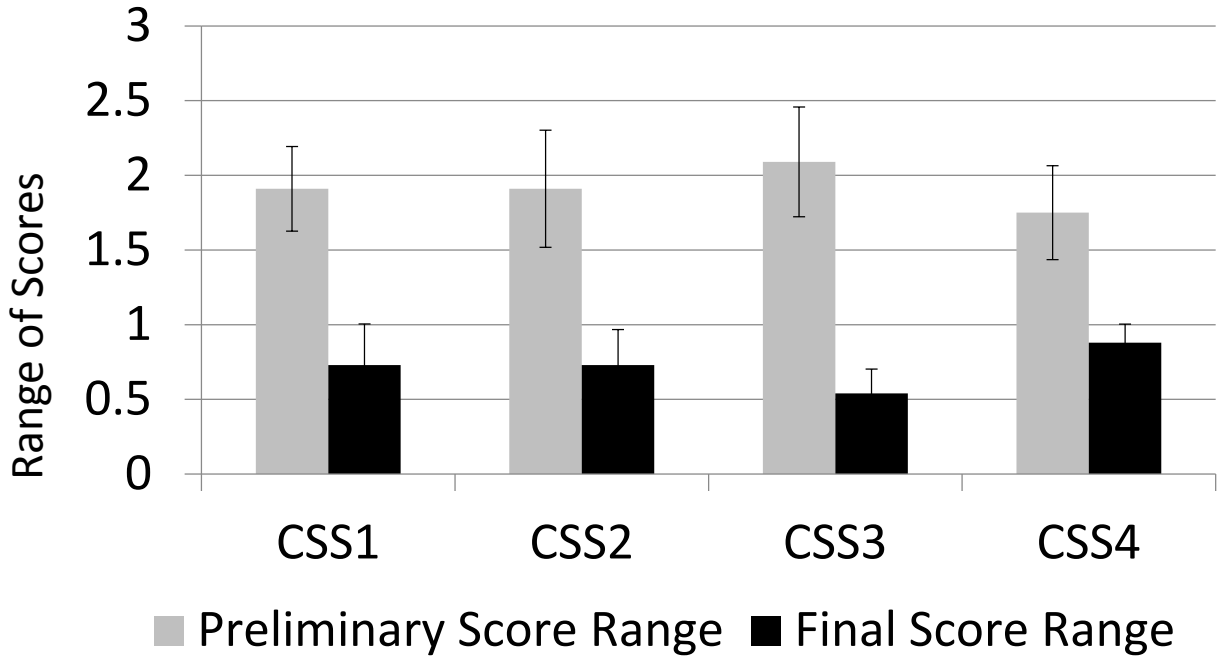
Worse agreement ***between*** panels after collaborative discussion

α = -.0517


Krippendorff's α - All Applications
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
Preliminary Reviewer Scores
Final Reviewer Scores
Final Panel Scores
Intra-Panel
Inter-Panel
Pier et al, 2017

Scoring Variability

How variable are reviewers' scores?



Low agreement among individual reviewers
α = .0840

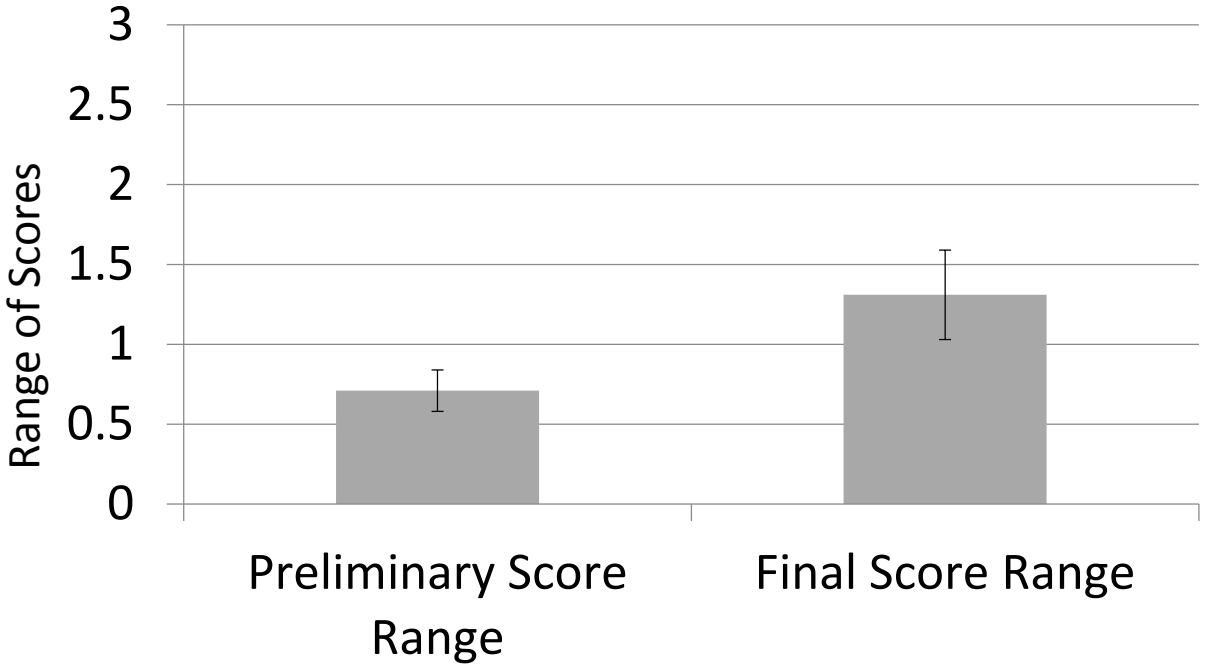
Better agreement *within* panels after collaborative discussion
α = .6652

Worse agreement *between* panels after collaborative discussion
α = -.0517

Pier et al, 2017

Scoring Variability

How variable are reviewers' scores?



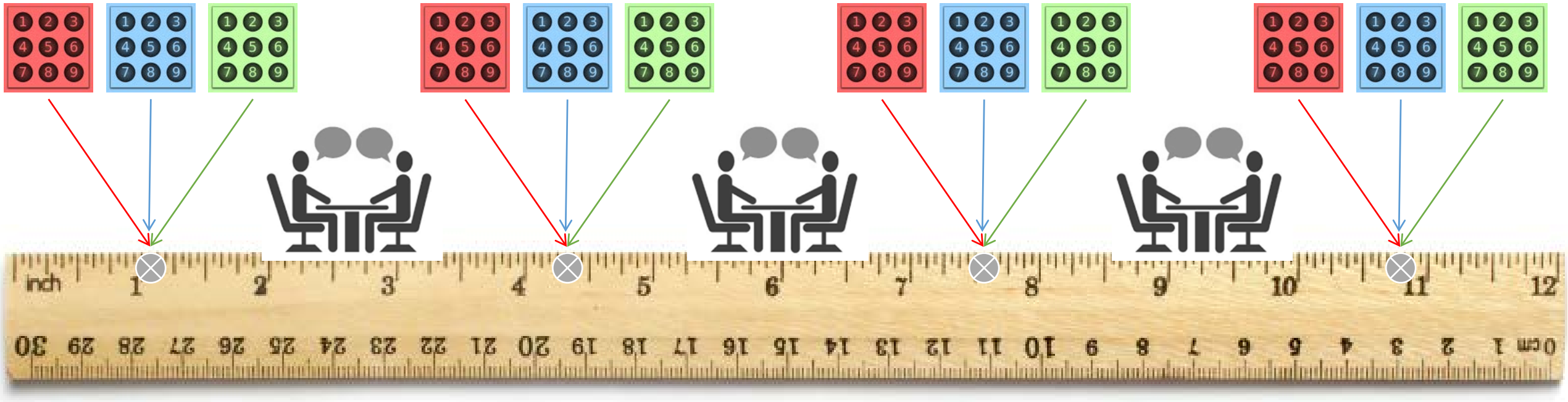
Low agreement among individual reviewers
α = .0840

Better agreement *within* panels after collaborative discussion
α = .6652

Worse agreement ***between*** panels after collaborative discussion
α = -.0517

Pier et al, 2017

Scoring Variability



Score Calibration Talk [SCT]

Overall Impact	*Score*	*Descriptor*
High	1	Exceptional
	2	Outstanding
	3	Excellent
Medium	4	Very Good
	5	Good
	6	Satisfactory
Low	7	Fair
	8	Marginal
	9	Poor

LZ-2: So that means uh probably they are already recognize this

JA-3: you highlight those differences you know or uh y'know

Score Calibration Talk [SCT]

Instances of Score Calibration Talk (SCT) in each CSS

	CSS1	CSS2	CSS3	CSS4	Total
Self-initiated SCT					
# Instances	15	18	11	12	56
Time (m:s)	3:33	4:36	2:09	2:37	12:55
Other-initiated SCT					
# Instances	7	3	4	1	15
Time (m:s)	6:07	4:28	5:27	1:46	17:48
Total SCT					
# Instances	22	21	15	15	71
Time (m:s)	9:40	9:04	7:36	4:23	30:43

Pier et al, 2017

Score Calibration Talk [SCT]

SCT & Reviewer Score Change

	Correlation
Self-initiated SCT	
# Instances	*r* = .108
Time (m:s)	*r* = .067
Other-initiated SCT	
# Instances	*r* = .978
Time (m:s)	*r* = .961
Total SCT	
# Instances	*r* = .717
Time (m:s)	*r* = .809

SCT & Reviewer Score Convergence

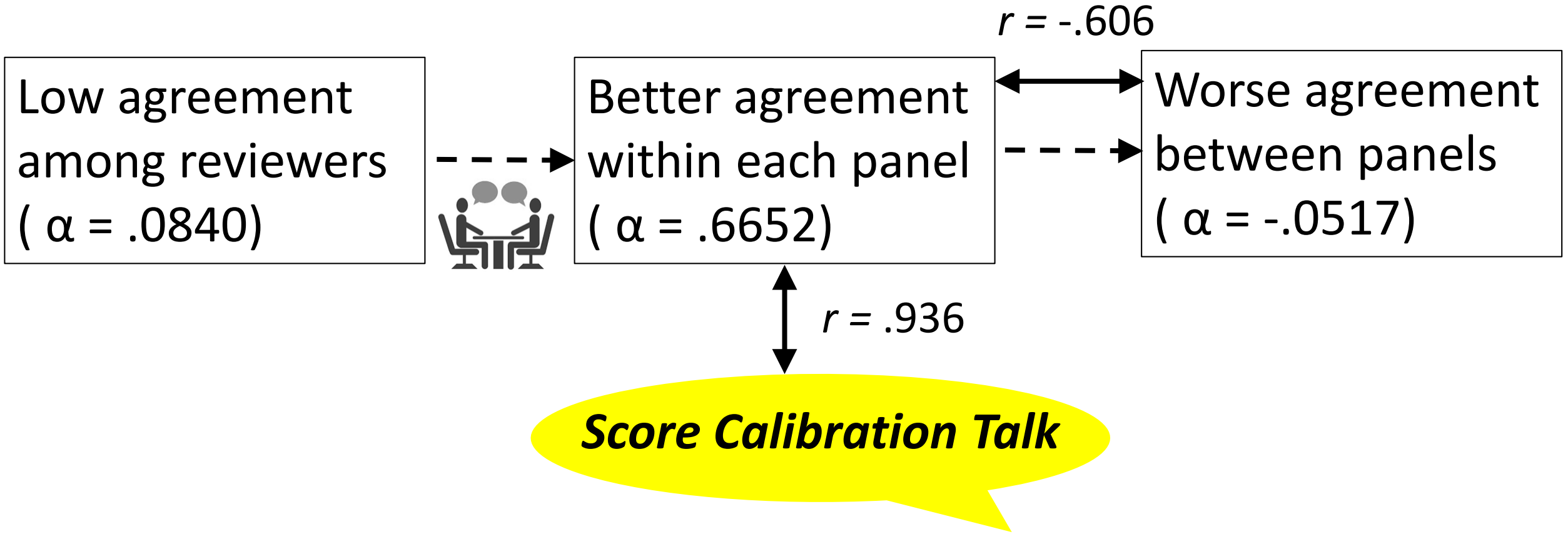
	Correlation
Self-initiated SCT	
# Instances	*r* = .682
Time (m:s)	*r* = .657
Other-initiated SCT	
# Instances	*r* = .858
Time (m:s)	*r* = .784
Total SCT	
# Instances	*r* = .980
Time (m:s)	*r* = .936

Within-panel convergence & Between-panel divergence

$r = -.606$ ($p = .005$)

Pier et al, 2017

Conclusions



Implications

- Identifying the interpersonal and communicative processes that unfold during peer review meetings helps us better understand how subjectivity can bias people's decision making

- SCT may be a prime target for intervention to help continue to improve the reliability of the peer review process

- Training SROs and Chairs about the local norming that happens during SCT could help them ensure more consistent adherence to the scoring rubric

Thank you!

CENTER FOR WOMEN'S HEALTH RESEARCH
University of Wisconsin-Madison

mlcarnes@wisc.edu

epier@wisc.edu

This work is supported by the National Institute of General Medical Sciences of the National Institutes of Health (Award # R01GM111002) and by the Arvil S. Barr Graduate Fellowship.