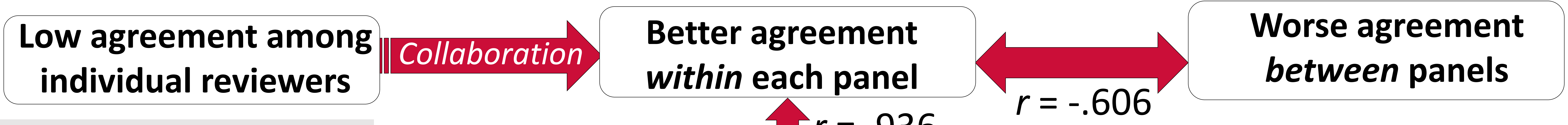
The Fallibility of Expert Scientists:
How *Score Calibration Talk* Undermines Fairness in the Scientific Peer Review Process

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Elizabeth L. Pier, Joshua Raclaw, Anna Kaatz,
Markus Brauer, Molly Carnes, Mitchell J. Nathan, & Cecilia E. Ford

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



Low agreement among individual reviewers

Collaboration

Better agreement *within* each panel

r = -.606

Worse agreement *between* panels

r = .936

Score Calibration Talk

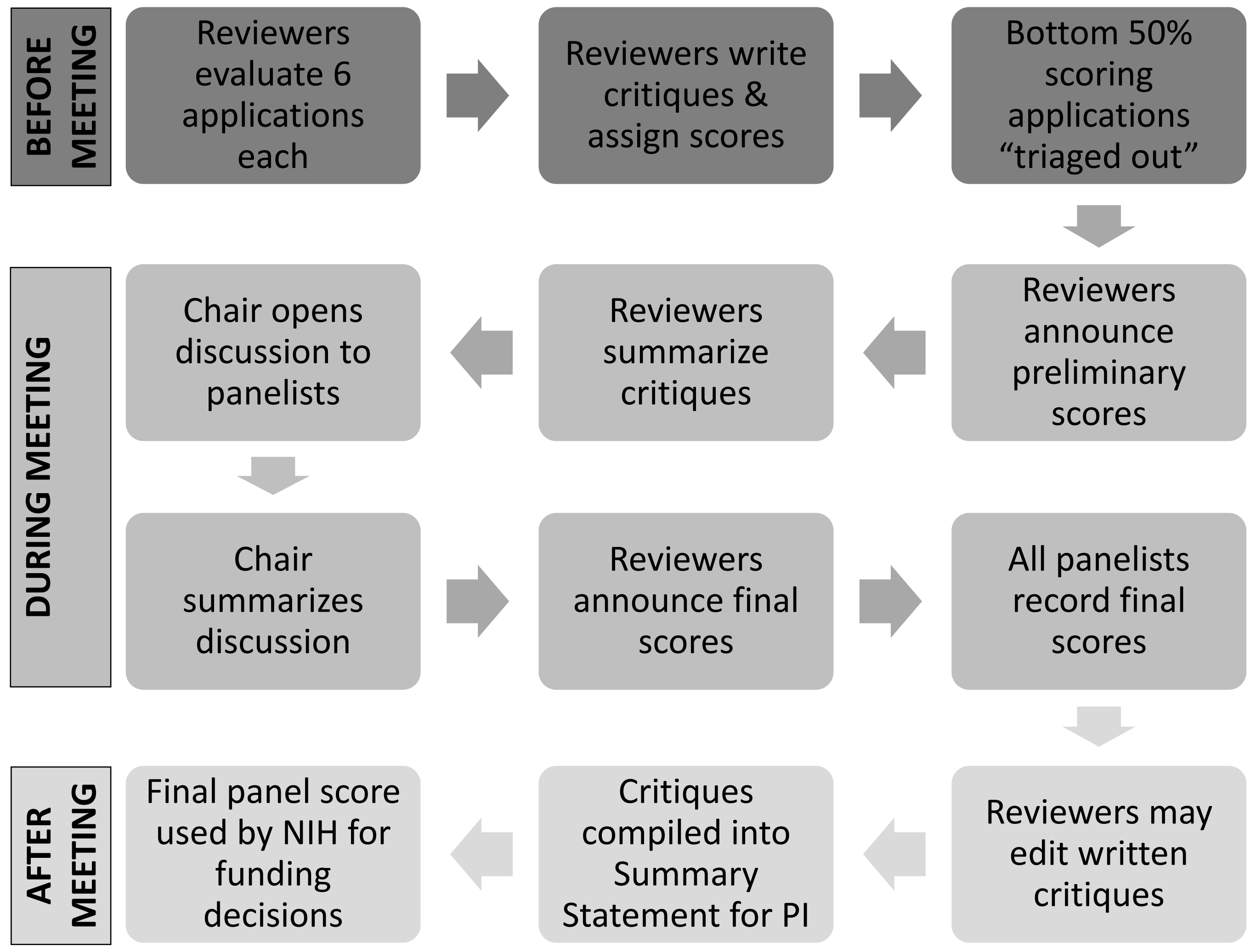
Constructed Study Sections

42 experienced reviewers for NIH participating in one of four videotaped ***Constructed Study Sections*** (CSS)

reviewed

25 R01 grant applications submitted between 2012 – 2015 to the Oncology 1 or Oncology 2 review groups within NIH's National Cancer Institute

De-identified screenshot from one CSS



BEFORE MEETING

Reviewers evaluate 6 applications each → Reviewers write critiques & assign scores → Bottom 50% scoring applications "triaged out"

DURING MEETING

Reviewers announce preliminary scores → Reviewers summarize critiques → Chair opens discussion to panelists → Chair summarizes discussion → Reviewers announce final scores → All panelists record final scores

AFTER MEETING

Reviewers may edit written critiques → Critiques compiled into Summary Statement for PI → Final panel score used by NIH for funding decisions

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

The research reported here was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM111002. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Self-Initiated SCT

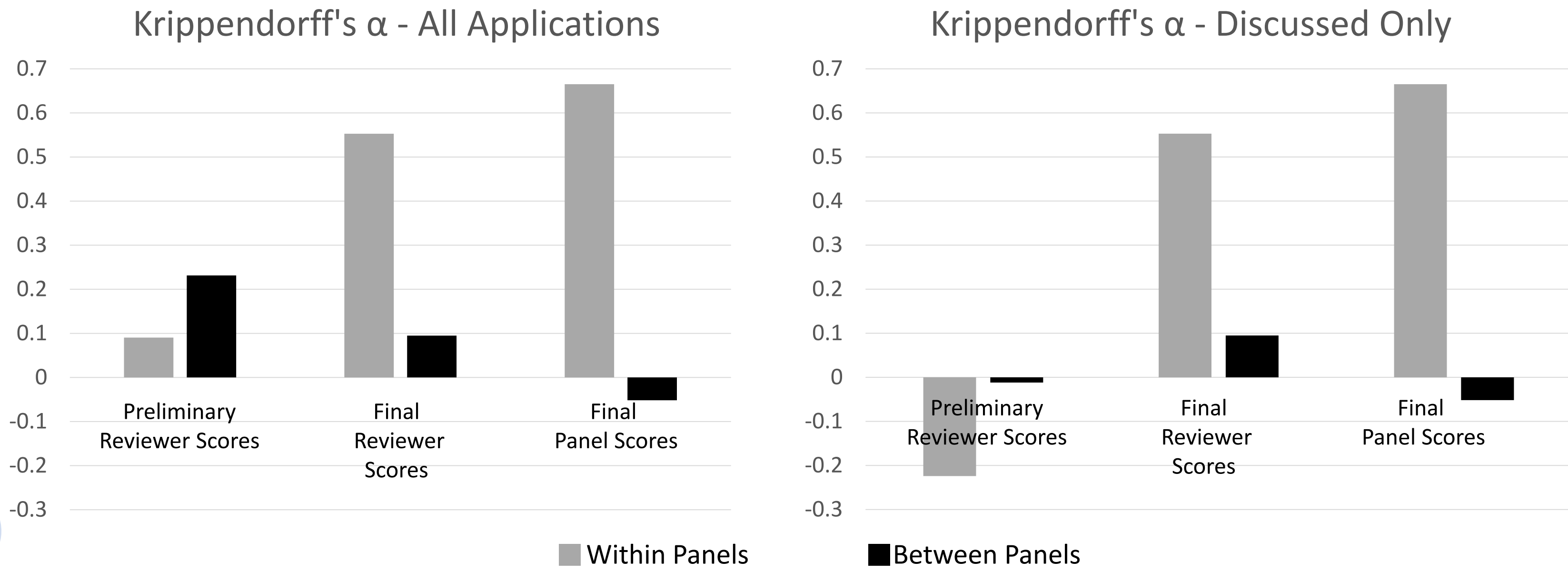
TB-2: Yeah so I gave it a one, and you know, as you mentioned before, you only give a one once in a lifetime, so to speak. And I thought that this was one of the the best grants I guess I've ever written—I've ever read, because really cause of three things. There is, I thought that the impact was large and obvious, and it was largely driven by quite a bit of of preliminary data...

Other-Initiated SCT

Chair: Other comments? *(pause)* So with that, let's hear our new scores?
MP-1: So I'll move to a four.
Chair: Secondary?
CV-2: Uh, I'll move to four also.
Chair: Dr. Joshi?
GJ-3: I had four to begin with and I'll stay there.
Chair: Anyone outside that ra—these are pretty serious concerns that were raised. Four is a very high score.
JR: Yeah.
CV-2: Yeah mine, actually go to a five. *(group laughter)*
Chair: Okay.
GJ-3: I'll go to five.
MP01: I'll go to five.
Chair: Let's go again. The preliminary-um new scores are? *(group laughter)* Preliminary? Dr. Patil?
MP-1: Five.
CV-2: Five.
GJ-3: Five.

Score Calibration Talk (SCT)

	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



Values of α > .80 are "reliable", .67 - .80 are "tentative" (Krippendorff, 2013)

Range of scores for an application significantly ***decreased*** within each panel after discussion:

	Prelim Range	**Final Range**	$t_{(df)}$, p
CSS1	*M* = 1.91 (*SD* = 0.94)	*M* = 0.73 (*SD* = 0.91)	t_{10} =3.99 *p* = .003
CSS2	*M* = 1.91 (*SD* = 1.30)	*M* = 0.73 (*SD* = 0.786)	t_{10} =4.49 *p* = .001
CSS3	*M* = 2.09 (*SD* = 1.22)	*M* = 1.09 (*SD* = 0.54)	t_{10} =2.80 *p* = .019
CSS4	*M* = 1.75 (*SD* = 0.89)	*M* = 0.88 (*SD* = 0.35)	t_{7} =2.97 *p* = .021

Range of scores for an application significantly ***increased*** between panels after discussion:

Prelim Range	**Final Range**	$t_{(df)}$, p
M = 0.71 (*SD* = 0.45)	*M* = 1.31 (*SD* = 0.97)	t_{11} =-2.19 *p* = .05

SCT & Scoring Variability

SCT &
Reviewer Score Change:

Self-Initiated SCT	**Correlation**
# 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 &
Panel Score Convergence:

Self-Initiated SCT	**Correlation**
# 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

Relationship between within-panel score converge & between-panel score divergence:

***r* = –.606** (*p* = .005)