A Kuhnian Perspective on Publication Biases and Low Inter-Reviewer Reliabilities

Carole J. Lee
University of Washington, Seattle
Center for Education and Social Science Research, Workshop in Methods
Indiana University
February 25, 2011

Empirical study of peer review

• Hypothesis-testing research on peer review (Fletcher & Fletcher 1997).

• Growing industry of empirical work, especially in psychology and medicine (Weller 2001).
Peer review as “key event in the history of modern science” (Ziman 1969).

Empirical research on peer review’s biases:
1. Low inter-rater reliabilities  
2. Publication biases

Kuhnian reinterpretation

The tension

The tension

Peer review as “key event in the history of modern science” (Ziman 1969).

Empirical research on peer review’s biases:
1. Low inter-rater reliabilities  
2. Publication biases

Kuhnian reinterpretation

reconceptualizes social epistemology
1. Empirical research on inter-rater reliability

Single-rater reliability = correlation between ratings for two reviewers on a single submission across many submissions.

<table>
<thead>
<tr>
<th>Single-rater reliability</th>
<th>Reliability of whom?</th>
<th>Study / Number of reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.18 to 0.37</td>
<td>Reviewers for NSF grants in chemical dynamics, solid-state physics, and economics</td>
<td>Cole et. al. 1981; Cicchetti 1991 / N = 50 proposals per field with 12 reviewers each</td>
</tr>
<tr>
<td>0.21</td>
<td>Reviewers for Australian Research Council grants in social sciences and humanities</td>
<td>Jayasinghe et. al. 2003 / N = 1928 reviewers</td>
</tr>
<tr>
<td>0.19</td>
<td>Reviewers for Australian Research Council grants in physical sciences</td>
<td>Jayasinghe et. al. 2003 / N = 4295 reviewers</td>
</tr>
<tr>
<td>0.17 to 0.28</td>
<td>Reviewers for papers submitted to <em>Am Soc R, Physiological Zoology, Law and Soc R, Am Psych, Per and Soc Psych Bull</em></td>
<td>Hargens &amp; Herting 1990 / N &gt;140 per journal</td>
</tr>
</tbody>
</table>
2. Psychometric assumptions about inter-rater reliability

A. Submissions have a latent overall quality value along a single dimension of evaluation.
B. Role of expert reviewers = reliably identify the latent quality value of submission (Hargens & Herting 1990).
C. How high should inter-rater reliability be?

<table>
<thead>
<tr>
<th>Inter-rater reliability</th>
<th>Psychometric tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 0.9</td>
<td>Intelligence tests</td>
</tr>
<tr>
<td>&gt; 0.7</td>
<td>Personality tests</td>
</tr>
<tr>
<td>~ 0.6</td>
<td>Essay marking</td>
</tr>
<tr>
<td>~ 0.2</td>
<td>Rorschach inkblot test</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Single-rater reliability</th>
<th>Criterion evaluated</th>
<th>Reliability of whom?</th>
<th>Study / Number of reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>Originality</td>
<td>Reviewers for Australian Research Council grants in social sciences, humanities, and physical sciences</td>
<td>Jayasinghe et. al. 2003</td>
</tr>
<tr>
<td>0.15</td>
<td>Methodology</td>
<td>Reviewers for Psychological Science Research Council grants in psychology and education</td>
<td>N = 6,223</td>
</tr>
<tr>
<td>0.16</td>
<td>Scientific/theoretical merit</td>
<td>Reviewers for Psychological Science Research Council grants in psychology and education</td>
<td>Marsh &amp; Ball 1989</td>
</tr>
<tr>
<td>0.12</td>
<td>Significance</td>
<td>Reviewers for <em>Educational Psychology</em></td>
<td>N = 778</td>
</tr>
<tr>
<td>0.23</td>
<td>Research design</td>
<td>Reviewers for <em>Educational Psychology</em></td>
<td></td>
</tr>
<tr>
<td>0.22</td>
<td>Clarity of problem, hypothesis, assumptions</td>
<td>Reviewers for <em>Journal of Personality and Social Psychology</em></td>
<td>Scott 1974</td>
</tr>
<tr>
<td>0.28</td>
<td>Importance of contribution</td>
<td>Reviewers for <em>Journal of Personality and Social Psychology</em></td>
<td>N &gt;200</td>
</tr>
<tr>
<td>0.19</td>
<td>Design and analysis</td>
<td>Reviewers for <em>Journal of Personality and Social Psychology</em></td>
<td></td>
</tr>
</tbody>
</table>
Peer review is a poor evaluation tool by Am Psych A's Standards for psychological testing (Bornstein 1991).

Inter-rater reliabilities are considered poor by psychometric standards (Suls 2009).

The critical direction for future research is how to improve the reliability of peer reviews (Jayasinghe et. al. 2003).
3. Reasonable disagreement

Reflexive felicity of psychometric research on peer review.

Inter-rater disagreement can reflect normatively appropriate differences in:
A. Sub-specialization
B. Interpretation/application of evaluative criteria

3. Reasonable disagreement

Case A: Differences in sub-specialization:
– Expertise in theory vs.
– Statistics, modeling, and/or sampling techniques vs.
– Domain-specific application of those techniques.

In cases where quality along different aspects diverge, we would not expect high inter-rater reliability scores (Hargens & Herting 1990).
3. Reasonable disagreement

Case B: Reviewers interpret/apply evaluative criteria differently.
   – Deeper disagreement in disciplinary cases?
     • Kuhn: interpretation and application of epistemic values like “accuracy” is underdetermined.
     • Same problem of underdetermination for evaluative criteria like “novelty” in peer review.

4. New empirical and philosophical questions

Empirical hypothesis: Experts can have diverging evaluations because they interpret or apply evaluative criteria differently.
4. New empirical and philosophical questions

Normative question: Are disagreements about the interpretation or application of evaluative terms normatively appropriate?

– Revolutionary Science: Mixed cases

In cases where reviewers chosen from different camps, we would not expect high inter-rater agreement.
5. Luck of the reviewer draw

Even if considerations raised by reviewers are not arbitrary, peer review outcomes can become an arbitrary result of which reviewer perspectives are brought to bear.

Outcomes:
- based “substantially on chance” (Jayasinghe et. al. 2003)
- result of “luck of the reviewer draw” (Cole et. al. 1981)

5. Luck of the reviewer draw

How should discipline-wide publication venues be structured to accommodate these problems?

- Number of prestigious disciplinary outlets available
- Thresholds (rejection rates) (Hargens & Herting 1990)
I. Inter-rater reliability: conclusions

“Essential aspects of the process generally known as verification will be understood only by recourse to the features with respect to which” researchers “may differ while still remaining scientists” (Kuhn 1977).

Essential aspects of peer review should be understood by recourse to the features in which reviewers may differ while still remaining experts in their field?

II. PUBLICATION BIAS

Different interpretations of novelty across disciplines.

Q: Why these differences from a Kuhnian point of view?

The answer will challenge how publication biases are categorized and measured.

Problem: Missing data for meta-analysis and future experiments.

Solution: Change incentives and communication structures.
1. Heterogeneity of categorized “bias”

<table>
<thead>
<tr>
<th>Study</th>
<th># Positive Trials Published / # Trials</th>
<th># Negative Trials Published / # Trials</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Bardy 1998)</td>
<td>52 / 111 (47%)</td>
<td>16 / 77 (53%)</td>
<td>(Bardy 1998)</td>
</tr>
<tr>
<td>(Dickersin et al, 1992)</td>
<td>84 / 96 (87%)</td>
<td>52 / 72 (72%)</td>
<td>(Dickersin et al, 1992)</td>
</tr>
<tr>
<td>(Dickersin &amp; Min 1993)</td>
<td>121 / 124 (98%)</td>
<td>63 / 74 (85%)</td>
<td>(Dickersin &amp; Min 1993)</td>
</tr>
<tr>
<td>(Ioannidis 1998)</td>
<td>20 / 27 (74%)</td>
<td>16 / 39 (41%)</td>
<td>(Ioannidis 1998)</td>
</tr>
<tr>
<td>(Stern 1997)</td>
<td>55 / 76 (72%)</td>
<td>18 / 54 (20%)</td>
<td>(Stern 1997)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>332 / 434 (76%)</td>
<td>165 / 316 (52%)</td>
<td><strong>Total</strong></td>
</tr>
</tbody>
</table>

2. Kuhn: new empirical and philosophical questions

“Novelty”
2. Kuhn: new empirical and philosophical questions

“Novelty”

Goals of medicine: to cure disease and ameliorate patient suffering (Horrobin 1990).

“Novel” = insofar as achieves these goals.

Prefer publishing trials showing higher efficacy than standard treatment (not placebo).

All things being equal, therapeutic considerations motivate a preference for:

1. Trials for new therapies showing a statistically significant advantage over standard treatment alone.

2. Trials for new therapies showing a statistically significant advantage over placebos.
2. Kuhn: new empirical and philosophical questions

Treatments shown to be ineffective when compared to placebo are therapeutically significant in some cases.

Important when demonstrate that current standards of care are not adequate.

Examples:
• Inefficaciousness of drugs for depression (Kirsch et al, 2002).
• Inefficaciousness of hormone replacement therapy to prevent heart disease in menopausal women (Hully et al, 1998).

2. Kuhn: new empirical and philosophical questions

All things being equal, therapeutic considerations motivate a preference for:

1. Trials for *standard* treatments shown to be inefficacious compared to placebo.

    over

2. Trials for *new* treatments shown to be inefficacious compared to placebos.
3. Problem of missing data

Skewed samples for meta-analyses.

Skews result in direction of effectiveness for new drugs.

Inefficiency for future researchers spending time and money repeating unsuccessful experiments.

4. Discipline-wide incentives and communication

Create registries that make all trial results public.

World Health Organization establishing International Clinical Trials Registry Platform.

However: Incentive for authors?
• Incentives to develop and submit “publishable” work.
III. CONCLUSION
III. CONCLUSION

Practical hurdles to publication bias research:

Access to journal reviews harder than access to author surveys.

Journal reviews needed to evaluate publication preferences.

thank you !