Evaluating the Reproducibility of Computer Searches Used for Systematic Reviews (SR)

Paul Fehrmann, Kent State University - Kent Campus

Available at: https://works.bepress.com/paul_fehrmann/5/
Evaluating the Reproducibility of Computer Searches Used for Systematic Reviews (SR)

“The search process needs to be documented in enough detail ... to the extent that all the searches of all the databases are reproducible.”

(Searching for studies: A guide to information retrieval for Campbell Systematic Reviews, Campbell Collaboration, 2010)

SR in Psychology
EBSCO’s PsycINFO. May 16, 2013
Searched English journal articles, “methods”=SR

Year  Hits  
2000  17  
2005  212  
2010  995  
2013  1,319  
The use of SR is growing.

Resources for Evaluating SR
• AMSTAR
• STARLITE
• PRISMA
• PRESS

A Computer Search Report Checklist (CSRC)

• To assess computer search steps reported in SR.
• For comprehensiveness (C)
• For reproducibility (R)

Why reproducible computer searches ?
• Increased confidence in SR results.
• Confident evaluation of SR search methods.
• Systematic extension of SR search methods.

Computer Search Report Checklist (CSRC)
Timeline, development, use, publication, ongoing work
2007/08
Reviewed major guidelines/recommendations for item content
Non-redundant recommendations translated into 47 items
2008/09
Trained and piloted test items
Assessed checklist item inter-rater reliability
2009/11
Evaluated psychology SR with "comprehensiveness items" (C)
Article published in journal Research Synthesis Methods
2012/present
Use "reproducibility”(R) items with sample of psychology SR
Results to be submitted for publication

Ongoing development
Gather input from experts
Target, refine, reduce number of items
Possible weighted scoring for R and C items
Possible web site to support CSRC scoring & reports

Inter-Rater Agreement & Frequencies for select Reproducibility Items

<table>
<thead>
<tr>
<th>CSRC Item</th>
<th>Frequencies***</th>
<th>Campbell</th>
<th>Psych</th>
<th>Fisher’s***</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Copy (any)</td>
<td>Kappa* = .951 (.000)</td>
<td>13/25</td>
<td>1/25</td>
<td>.000</td>
</tr>
<tr>
<td>2. Names (every)</td>
<td>Kappa* = .485 (.000)</td>
<td>25/25</td>
<td>24/25</td>
<td>1.000 (ns)</td>
</tr>
<tr>
<td>3. Vendor (any)</td>
<td>Kappa* = .896 (.000)</td>
<td>10/25</td>
<td>2/25</td>
<td>.018</td>
</tr>
<tr>
<td>5. Date run (any)</td>
<td>Kappa* = .700 (.000)</td>
<td>10/35</td>
<td>3/25</td>
<td>.051</td>
</tr>
<tr>
<td>7. All exact terms (any)</td>
<td>Kappa* = .291 (.031)</td>
<td>22/25</td>
<td>24/25</td>
<td>.609 (ns)</td>
</tr>
<tr>
<td>13. Term combos (every)</td>
<td>Kappa* = .263 (.024)</td>
<td>22/25</td>
<td>18/25</td>
<td>.289 (ns)</td>
</tr>
<tr>
<td>28. Two independent IE decisions at level of title/abstract (every)</td>
<td>Kappa* = .606 (.000)</td>
<td>13/25</td>
<td>3/25</td>
<td>.005</td>
</tr>
</tbody>
</table>

25 systematic reviews were randomly selected from PsycARTICLES and 25 were randomly selected from the Campbell Library. Select item results above.

* Cohen’s Kappa was used to assess inter-rater agreement for the checklist items. Many item Kappa values were significant at <.001. Kappa values below 40 have been interpreted as showing poor to fair agreement, those between .40 and .60 moderate agreement, those between .60 to .80 substantial agreement, and values above .80 excellent agreement (Portney & Watkins, 1993).

** Frequencies are the number of Campbell and PsycARTICLES SR that scored Y for an item after first collapsing scores to Y or N, and then reaching consensus.

*** Fisher’s Exact test was used as a measure of the significance of item score differences between the PsycARTICLES and Campbell Collaboration SR. Non-significant differences are noted “ns”.

References

Current CSRC information: libguides.library.kent.edu/CSRC

Paul Fehrman pfehrman@kent.edu
Matthew Cox mcox19@kent.edu