On the Future of Evaluation and BELIV

Tamara Munzner
Department of Computer Science
University of British Columbia

BELIV 2016 Panel
October 24 2016, Baltimore MD

http://www.cs.ubc.ca/~tmm/talks.html#beliv16panel

@tamaramunzner

I sense a great disturbance in the Force:

The replication crisis in psychology &
its possible repercussions for the future of
vis evaluation;

or, What's making me lose sleep lately

The replication crisis in psychology

- early rumblings left me with (ignorable) qualms
  - papers: Is most published research false?, Storks Deliver Babies (p= 0.008), The Earth is spherical (p < 0.05), False-Positive Psychology
- groundswell of change for what methods are considered legitimate
  - out
    - p-value fishing / data dredging
    - Hypothesizing After Results are Known (HARKing)
  - in
    - replication
    - pre-registration
  - brouhaha with bimodal responses
    - some people doubling down and defending previous work
    - many willing to repudiate (their own) earlier styles of working

Remarkable introspection on methods

- thoughtful willingness to change standards of field
  - Andrew Gelman's commentary on the Susan Fiske article
    - http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/
  - Simine Vazier's entire Sometimes I'm Wrong blog
    - http://sometimesimwrong.typepad.com/
    - especially posts on topic Scientific Integrity
  - Joe Simmons Data Colada blog post What I Want Our Field to Prioritize
    - http://datacolada.org/53/
  - Dana Carvey's brave statement on her previous power pose work
    - http://faculty.haas.berkeley.edu/dana_carney/pdf_My%20position%20on%20power%20poses.pdf

When and how will this storm hit us?

- they're ahead of us
  - they have some paper retractions
    - we don't (yet) have any retractions for methodological considerations
  - they agonize about difficulty of getting failure-to-replicate papers accepted
    - we hardly ever even try to do such work
  - they are a much older field
    - we're younger: might our power hierarchies thus be less entrenched??…
  - they are higher profile
    - we don't have vis research results appear regularly in major newspapers/magazines
  - they have rich fabric of blogs as major drivers of discussion
    - crosscutting traditional power hierarchies
    - we have far fewer active bloggers

Terrain of blog critiques

- meta: methods for methodological critique
  - Uri Simonsohn DataColada post on civility
    - http://datacolada.org/52
    - don't label, describe
    - don't infer motives
    - reach out: contacting authors whose work you discuss before making things public
      - as a heuristic check on tone, imagine going to dinner with authors and their parents that night
- resonates with my own first foray into blog critique
    - https://tamaramunzner.wordpress.com/2016/01/16/on-the-memorability-debate/
  - tone check advice is spot on
    - I *did* go out to dinner with Stephen Few the night I wrote my blog posts!
      - leading me to pick my tone with suitable care
  - I did not reach out, but now I think it would be wise indeed

Is BELIV still needed?

- yes yes yes! so much to think about!

- progression / cycle
  - no evaluation
  - first wave of quantitative evaluation
  - second wave of qualitative evaluation
  - now: third wave revisiting quantitative evaluation??