**PAPER NUMBER #402**

**A practitioner’s tool for critical appraisal of randomized assignment and quasi-experimental program evaluations**

**Presenting Author**

Annette N. Brown

**Affiliation**

FHI 360

**Country of residence**

United States of America

**Objectives/aims**

In an ideal situation, practitioners (policy analysts and program managers) are able to obtain the evidence they need from carefully conducted systematic reviews. In practice, however, the consumers of evidence often call on evidence from a single study or need to review evidence from multiple studies rapidly to make pressing decisions or respond to solicitations with deadlines. Most critical appraisal tools have been developed for systematic review teams with many researchers and much time. I designed a tool to help practitioners better understand the reliability of impact evaluations conducted by both public health and social science researchers.

**Methods**

The tool is based on Waddington et al.’s (2017) four categories of study design, which flow from randomized assignment to single-difference designs. The first step is to identify the study’s design and then the reviewer answers five or fewer questions adjust the score from there. To better assess evidence from social science, the tool accounts for whether the included study is a working paper. Another distinguishing feature is that the tool accounts for whether there is a publicly available internal replication of the study. The other scoring factors cover the standard domains of bias, but use intuitive questions instead of methodological questions.

**Main findings**

The scoring uses four levels but condenses the final appraisals into three categories – few limitations, some limitations, and many limitations. I conduct a validation test by having two reviewers, one a program manager with a master’s degree and one a researcher with a Ph.D., independently appraise two sets of impact evaluations. I calculate inter-rater reliability for each set using the Kendall and Spearman statistics for correlation coefficients between two raters when the rating scale is ordinal. I compare the findings to reliability statistics for other critical appraisal tools, specifically those that appraise non-random assignment studies.