Reviewer's report

Title: Exploring inter-rater reliability and measurement properties of environmental ratings using kappa and colocation quotients

Version: 1
Date: 11 July 2014

Reviewer: Timothy Leslie

Reviewer's report:

I have read ‘Exploring inter-rater reliability and measurement properties of environmental ratings using kappa and colocation quotients’ submitted for consideration in Environmental Health. My background relevant to this review is that I am the author of the statistic being investigated (the CLQ).

The manuscript is an assessment of how a newer statistic (the CLQ) compares with a very widely used standard (kappa) in assessing inter-observer reliability. As a case study for this test, a survey of ‘greenness’ was conducted in Sweden. The paper is well written and clearly organized. The tables are generally well organized, although I did not find the figures appreciably useful. It was nice to see the statistics were sufficiently powered, with almost 28,000. Overall I believe the paper merits publication subsequent to revisions in two notable veins:

1) Message: The strongest conclusion of this paper as written is that the CLQ and the kappa are equally useful. This is listed as a relatively unsurprising outcome, and given that the CLQ is decomposed into a pure odds-ratio statistic, perhaps it is not. However, there is little discussion of the situations where kappa or the CLQ might be more useful (or what information this new statistic can contribute). The CLQ is noted as being spatial in nature and not having a fixed maximum value, but no real advantages to either statistic are discussed at any length: they are noted as having quite similar properties.

2) Category Collapse: We would hope for inter-observer reliability, the finding that there was co-location between the two disagree categories and the two agree categories was expected. That disagreement and agreement do not share much spatial overlap is good to see for robustness. Using the kappa and CLQ as a guide for collapsing categories is an interesting choice, but is done on a somewhat ad hoc basis (what sort of cutoffs would the authors deem suggestive of category collapse?). The collapsed groupings have new disagreements that were not present (the agreements with the cannot say / not answered group) in the original set. In the second example, while the authors discuss how the reliability similarities differ from the first example, the authors do not investigate collapsing the disagree categories while leaving the agree categories (which don’t share the overlap of the first example), or state why collapsing categories in such a situation is inappropriate or uninteresting.

Level of interest: An article of importance in its field
Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:
I declare that I have no competing interests