Reply to Comment on 'Quantifying the consensus on anthropogenic global warming in the scientific literature'

2015 Environ. Res. Lett. 10 039002

(http://iopscience.iop.org/1748-9326/10/3/039002)

Environmental Research Letters

OPEN ACCESS

RECEIVED
12 October 2014

ACCEPTED FOR PUBLICATION
31 January 2015

PUBLISHED
19 March 2015

REPLY

Reply to Comment on 'Quantifying the consensus on anthropogenic global warming in the scientific literature'

John Cook[1,2,3] **and Kevin Cowtan**[4]

[1] Global Change Institute, University of Queensland, Australia
[2] Skeptical Science, Brisbane, Queensland, Australia
[3] School of Psychology, University of Western Australia, Australia
[4] Department of Chemistry, University of York, UK

E-mail: j.cook3@uq.edu.au

Keywords: consensus, global warming, climate change

Inter-rater reliability statistics may be trivially calculated from the released data available at www.skepticalscience.com/docs/tcp_allratings.txt. We have placed R-code for this calculation on the project website at www.skepticalscience.com/docs/interrating.r. The unweighted Cohen kappa is 0.35 using the seven fine-grained categories used in the initial rating process. However, the consensus statistics are based on only three categories: 'endorse', 'reject' or 'no position'; for these categories, kappa rises to 0.46. Subdividing rating categories is known to depress kappa values. The more appropriate Fleiss kappa gives the same results. In our view, the categories should be considered as nominative (Cook *et al* 2014). However, if they are treated as ordered, the kappa value for the fine-grained categories approaches the value for the consensus categories. Kappa values are also depressed in the case when category counts are very uneven (Sim and Wright 2005). Our data is an extreme case with two orders of magnitude difference between the most and least populous categories.

The interpretation of these statistics is problematic. Landis and Koch (1977) propose an ad-hoc metric by which the agreement on the fine-grained categories would be called 'fair' and on the consensus categories 'moderate'. However, there is no theoretical basis for these labels. Dean cites Kottner *et al* (2011), who discuss kappa values for a rather different application (medical diagnosis), in which the accuracy of *individual* ratings has consequences for patient health. If however the physician were simply conducting a survey of the *prevalence* of a condition, agreement rates are less critical as long as the ratings are not biased. Similarly in our case, the agreement rate affects the uncertainty in the result, but only a bias would lead to an incorrect value for the consensus.

Because the consensus ratio is determined by two of the three categories, differences in allocation of papers to the 'no position' category have minimal impact on the conclusions. The proportion of ratings in the relevant categories (i.e. endorse, no position, reject) for the 12 raters who contributed at least 500 ratings were decomposed by change of variable into consensus invariant and consensus altering terms. The inter-rater variability in the consensus invariant variable was more than twenty times larger than in the consensus altering variable. Thus the primary cause of inter-rater variability arises from differing interpretations of the no-position criteria, but at the same time the raters applied their individual criteria consistently to both the endorse and reject categories. This suggests that inter-rater variability could be substantially reduced by clarification and training on the no-position criteria, but that doing so would not affect the final consensus percentages.

The final consensus percentages calculated for the 12 most prolific raters gives an estimate of the uncertainty in the results. Extreme values were 95.7% and 98.2%, with an interquartile range of 96.2% to 97.6%.

Potential bias among the raters was tested a second way by use of the author self-ratings (bearing in mind that the authors had access to the whole paper). The author ratings were assumed to be correct and were then used to calculate a correction to the abstract ratings. This correction was then applied across all the abstracts, to estimate the consensus score which would have been obtained had the authors rated all of the papers. The results are virtually unchanged (97.2% versus 97.1%). Thus this second method of bias evaluation also suggests that bias was not a significant problem. Nonetheless, we encourage third parties to independently examine the abstracts as a further audit of our results. Tools have been made available to facilitate this task at www.skepticalscience.com/tcp.php.

References

Cook J, Nuccitelli D, Skuce A, Way R, Jacobs P, Painting R, Honeycutt R, Green S A, Lewandowsky S and Coulter A 2014 *24 Critical Errors in Tol (2014): Reaffirming the 97% Consensus on Anthropogenic Global Warming* (http://sks.to/24errors)

Kottner J, Audigé L, Brorson S, Donner A, Gajewski B J, Hróbjartsson A, Roberts C, Shoukri M and Streiner D L 2011 Guidelines for reporting reliability and agreement studies (GRRAS) were proposed *J. Clin. Epidemiol.* **64** 96–106

Landis J R and Koch G G 1977 The measurement of observer agreement for categorical data *Biometrics* **33** 159–74

Sim J and Wright C C 2005 The kappa statistic in reliability studies: use, interpretation, and sample size requirements *Phys. Ther.* **85** 257–68