QUANTIFYING THE CONSENSUS ON ANTHROPOGENIC GLOBAL WARMING IN THE LITERATURE: A COMMENT

Richard S.J. Tol

Department of the Economics, University of Sussex, Falmer, United Kingdom

Institute for Environmental Studies, Vrije Universiteit, Amsterdam, The Netherlands

Department of Spatial Economics, Vrije Universiteit, Amsterdam, The Netherlands

Tinbergen Institute, Amsterdam, The Netherlands

9 June 2013

Introduction

In a recent paper published in this journal (Cook et al. 2013), it was argued that 97% of the published literature on climate change supports the position that climate change is real and largely human-made. The paper was highlighted on the journal's website and picked up by the media around the world.

It is a strange claim to make. Consensus or near-consensus is not a scientific argument. Indeed, the heroes in the history of science are those who challenged the prevailing consensus and convincingly demonstrated that everyone else thought wrong.

A claim of consensus serves a political purpose, rather than a scientific one. This is fine. A consensus claim that climate change is really human-made is presumably made in support of the argument that greenhouse gas emissions be reduced. The view that emissions should be cut is widely shared, but not universal. Those who oppose climate policy do so for various reasons. Some argue that the impacts of climate change are no reason for concern, or that other problems take priority. Others argue that climate policy is ineffective and expensive. An argument that climate change is really human-made does not affect those positions. There are those who think that consensus is a sign of conspiracy. Some oppose climate policy out of self-interest. They may express concern about climate research to hide their motives, but they will not be swayed by new evidence.

Others, however, are concerned about the standards of proof in climate research. They would emphasize the complexities of the climate system and highlight lack of rigour in peer-review, substandard statistical analysis, and unwillingness to share data. These people are unlikely to be convinced by Cook et al. It is well-known that most papers and most authors in the climate literature support the hypothesis of anthropogenic climate change. It does not matter whether the exact number is 90% or 99.9%. These people are concerned about the quality of the research. More papers does not mean better papers.

In fact, the paper by Cook et al. may strengthen the belief that all is not well in climate research. I argue below that data are hidden, that the conducted survey did not follow best practice, that there are signs of bias in the data, and that the sample is not representative. In sum, the conclusion of Cook et al. does not stand. It may well be right, but it does not stand.

The survey

12,465 abstracts were downloaded from the Web of Science. The query was "global climate change" or "global warming". Only articles in English published between 1991 and 2011 were included. After cleaning, 11,944 abstract remained. The abstracts were assessed by a team of 24 volunteers[1], recruited through *Skeptical Science*, a polemic blog on climate change. Abstracts were rated on a 1-7 scale ranging from explicit, quantified endorsement of the human contribution to climate change to an explicit, quantified rejection.

Part of the survey results are available online: year of publication, title, journal, authors, classification, and reconciled rating. I repeatedly requested more information – specifically, first rating, second rating, third rating (if applicable), fourth rating (if applicable), rater ID (1^{st}, 2^{nd}, 3^{rd}, 4^{th}), time of rating (1^{st}, 2^{nd}, 3^{rd}, 4^{th}), author rating, survey protocol, and lab notes – but in vain.[2] Cook et al. thus confirm the hypothesis that climate researchers are secretive.

Cook et al. claim that 97% of abstracts endorse the hypothesis of anthropogenic climate change. The available data, however, has 98%.

After three rounds of rating were completed, a sample of 1,000 of the 7,970 papers rated 4 were reassessed, and split into 4a and 4b. 40 papers were classified as 4b, taking the consensus level from 98% to 97%. It is unclear whether they found 40 in the sample of 1,000, or 5 and scaled it up to 40 for the 7,970 neutral abstract. If the former is true, then 319 should have been reclassified. The headline endorsement rate would be 91% in that case. No survey protocol was

[1] Although Cook et al. claim that the ratings were independent, discussions between the raters have surfaced: http://rankexploits.com/musings/2013/i-do-not-think-it-means-what-you-think-it-means/

[2] Cf. (Singer 2008), http://www.aapor.org/Best_Practices1.htm and http://www.amstat.org/committees/ethics/index.html

published,[3] so it is unclear whether the 4th rating was an ad hoc addition. Data for the 4th rating are not available. The headline conclusion is not reproducible.

Signs of bias

The results depend on the quality of the rating.[4] Are the volunteers neutral observers, or are they predisposed to endorsing or rejecting anthropogenic climate change? Did they suffer from fatigue after rating a certain number of abstracts? 12 volunteers rated on average 50 abstracts each, and another 12 volunteers rated an average of 1922 abstracts each. This level of effort by a volunteer could indicate a strong interest in the issue at hand.[5] Fatigue may well have a problem.[6]

The Web of Science presents papers in an order that is independent of the contents of the abstract: Papers are ordered first on the year of publication, and second on the date of entry into the database. Abstract were randomly reshuffled before being rated. The data provided are in a different order again.[7]

In the data provided, raters are not identified and time of rating is missing. I therefore cannot check for inconsistencies that may indicate fatigue. I nonetheless do so. Figures S1-S9 shows the 50-, 100- and 500-paper rolling standard deviation, first-order autocorrelation – tests for fatigue – and skewness – a test for drift. I bootstrapped the data 10,000 times to estimate the expected value of these indicators and the 95% confidence interval. Table 1 summarizes the exceedence frequencies.

The data do not behave as expected. Rolling standard deviations are occasionally too large, and more frequently so than would be expected by chance alone. This may be because, in part of the sample, raters alternated between strong endorsement (1) and strong rejection (7). It may also be because, in part of the sample, large numbers of abstracts were rated near the mean. First-order autocorrelation should be zero, but it is not. In parts of the sample, ratings are consistently above average – perhaps because long sequences of abstracts were rated neutral (4). Figures S4-6 show local, first-order autocorrelation. Figure S10 shows global, higher-order autocorrelation for the reported data (ordered by year first and title second) and the data in alphabetical order. There is positive autocorrelation – and stronger autocorrelation in the alphabetical data – suggesting that abstracts were rated neutral (4) on the basis of their title. Skewness should be constant throughout the sample, but it is not. There is drift (towards endorsement of anthropogenic

[3] Cf. (Mohler et al. 2008)
[4] Cf. (Lyberg and Biemer 2008)
[5] Although not mentioned by Cook et al., at least 3, probably 5 and perhaps 9 of the volunteers are authors of the paper: http://www.skepticalscience.com/pics/tcp_raters3.gif Disconcertingly, John Cook managed the survey while rating papers.
[6] Indeed, one of the raters, Andy S, worries about the "side-effect of reading hundreds of abstracts" on the quality of his ratings. See http://rankexploits.com/musings/2013/i-do-not-think-it-means-what-you-think-it-means/
[7] The data provided do not detail the order of rating.

climate change) in the first fifth of the sample. Some parts of the sample show more negative skew than would be expected by chance: Endorsements are clustered. It thus appears that rating was not done consistently, perhaps because the raters tired.

Every abstract was rated twice. The data provided have only one rating, so I cannot check for consistency. The original authors report "disagreement" on "33% of endorsement ratings". About half of these were solved by a reconciliation process between the raters, while the other half were referred to a third rater. A comparison between the initial ratings and the final ratings would yield useful information on the validity of the ratings.

There are three duplicate records among the 11944 abstracts, and one case of self-plagiarism. This implies that there are four abstracts that are identical to another abstract. Of these four, two were rated differently – an error rate of 50%, after double rating and perhaps reconciliation or third rating.

The authors of the sampled papers were approached to rate their own work. This data could be used to validate the abstract ratings. Such a test is not reported in the original paper – only means are compared. A number of authors have come out to publicly state that their papers were rated wrong, but their number is too small for any firm conclusion. However, it is clear from Table 5 in Cook et al. (1) that the subsample of abstracts that were also rated by the authors is not representative for the whole sample ($\chi_2^2 = 22$; p<0.001) and (2) that the paper ratings are different from the abstract ratings ($\chi_2^2 = 5793$; p<0.001).

The majority of the selected papers are not on climate change itself, but rather on its impacts or on climate policy. The causes for climate change are irrelevant for its impact. Therefore, impact papers should be rated as neutral. Emission reduction policy would be pointless if climate change were not human-made, so policy papers can be rated as an implicit endorsement of the hypothesis of anthropogenic climate change. However, a paper discussing, say, carbon capture and storage cannot be taken as evidence for global warming. Even if the author firmly believes in human-made climate change and expresses that, she is an expert on carbon capture and storage and her opinions on the causes of climate change are irrelevant. These papers should therefore also be rated as neutral.

Table 2 shows the number of papers by rating and subject. Many papers that can only be rated as neutral in fact were not. In fact, 34.6% of papers that should have been rated as neutral were in fact rated as non-neutral. Of those misrated papers, 99.4% were rated as endorsements.

Table 3 shows the levels of endorsement, defined as the number of papers that support the hypothesis of anthropogenic climate change over all paper that take a position. For the whole sample, the ratio is 98.0% (out of 3974 papers). Counting only explicit endorsements (as implicit endorsements may be in the mind of the reader only), the ratio falls to 97.6% (out of 1010 papers).

Table 3 splits the sample into papers on impacts and mitigation (which have nothing to say on the causes of climate change) and papers on methods and palaeoclimate (which might have something to say on the causes of climate change). The endorsement level is much higher in impacts and mitigation (99.4% and 98.6%) than in methods and palaeoclimate (92.8% and 94.4%). The overall level of endorsement of the hypothesis of anthropogenic climate change is thus driven by papers that are not about the causes of climate change at all.

If I assume that methods and palaeoclimate papers are misrated in the same proportion as impacts and mitigation papers are, then the level of endorsement falls further, to 89.9% (implicit and explicit endorsement) and 93.8% (explicit endorsement only).

These levels of endorsement should be interpreted with care. If we take the numbers of Cook et al. at face value, 67% of the sampled papers did not take a position. If we move the papers on impacts and mitigation to 'neutral', 93% did not take a position. If we correct the relevant papers for misclassification, 95% of surveyed papers were silent on the hypothesis of anthropogenic global warming.

Trends

A key result by Cook et al. is that the level of endorsement of the hypothesis of anthropogenic climate change has increased over time. Figure 1 replicates this. There is indeed an upward trend ($p=0.052$). Figure 1 also shows the levels of endorsement in the abstracts classified as impacts and mitigation, and as methods and palaeoclimate. There is no upward trend in either ($p=0.249$ and $p=0.342$). The level of endorsement in impact and mitigation is much higher than in methods and palaeoclimate. The share of impact and mitigation in all abstracts has grown over time ($p=0.00003$). The apparent trend in endorsement is thus a trend in composition rather than in endorsement.

Representativeness

The sample includes almost 12,000 papers. The population of papers on climate change is much larger. Cook et al. do not test the representativeness of their sample. The sampling strategy rests on two crucial decisions: the data source, and the query. Cook et al. searched for papers on "global climate change" or "global warming". For the last 20 years or so, however, climate change has meant global climate change, unless otherwise specified. Replicating their query in May 2013, I find 13,458 papers. Their raw sample from May 2012 was 12,465. The additional thousand papers are probably later additions to the Web of Science. Dropping the "global" in "global climate change", I find 53,359 papers. That is, 74.8% of the population (or rather, a

larger sample) was excluded. There is nothing wrong with sampling, of course, as long as the sampling strategy leads to a representative sample (or sample weights to restore representation).

The Web of Science provides aggregate statistics for any query results. Figure 2 compares the disciplinary composition of the larger sample to that of the smaller sample. There are large differences. Particularly, the narrower query undersamples papers in meteorology (by 0.7%), geosciences (2.9%), physical geography (1.9%) and oceanography (0.4%), disciplines that are particularly relevant to the causes of climate change.

I next compare the output of the 100 most prolific authors in the bigger sample to their output in the smaller sample. Figure 3 shows the top 50. Many papers by some of the most active researchers were omitted. Although 25.2% of papers in the larger sample are in the smaller sample, only 20.0% of papers by the most prolific authors are included. This suggests that the narrower query undersampled the most active scholars.

Figure 4 shows the 50 most cited papers in the larger sample. Only 17 of those are included in the smaller sample, or 34%. This suggests that the narrower query oversampled the most influential papers.

The data behind Figure 2 suggests that the smaller sample included too many papers that are off-topic, which may have left more room for the volunteers to impose their preconceptions, that is, a bias towards the hypothesis of anthropogenic climate change. The data behind Figures 3 and 4 suggest that the smaller sample favoured influential authors and papers, who overwhelmingly support the hypothesis of anthropogenic climate change.

The data source is the other main decision in the sampling strategy. Cook et al. chose the Web of Science. I therefore posed the same queries to the Web of Science and to Scopus, a data source with similar functionality but wider coverage. Scopus returned 20,772 papers, 54% more than the Web of Science. Scopus uses fewer disciplines, so I aggregated the Web of Sciences disciplines to the Scopus ones. Figure 5 compares the results. As above, the disciplinary distribution of the smaller sample is not representative for the larger sample. In this case, earth and planetary sciences, the most relevant papers, are oversampled. This introduces a bias against endorsement.

Geophysical Research Letters is the most prominent journal in the query to both databases. However, Scopus returns 728 papers and the Web of Science 334. This is because the latter only considers the title, abstract and keywords, whereas the former uses meta-data too. Apparently, in more specialized journals, authors do not include a reference to “global climate change” or “global warming” but rather use more specific words. Scopus adds higher level keywords and thus retrieves such papers, whereas the Web of Science does not.

The Web of Science is more exclusive than Scopus. Young journals and obscure journals are better represented in Scopus. Such journals tend to be kinder on heterodox material. However,

this pro-establishment bias of the Web of Science is dominated by its omission of meta-data, which leads to the exclusion of more technical papers in more specialized journals.

Conclusion

In a recent paper, Cook et al. claimed that 97% of the literature endorsed the hypothesis that climate change is real and largely caused by human activity. Although they surveyed a large number of abstracts, most are not on the subject of the causes of climate change. Their consensus is not a consensus on the causes of climate change, but rather a vote of confidence of the broader climate research literature in the narrower climate science literature.

The reported data show signs of inconsistent rating, and a bias towards endorsement of the hypothesis of anthropogenic climate change. These concerns could easily be dismissed with the full data-set. Unfortunately, the authors have chosen not to make those available.

Although the number of surveyed papers is large, the number of papers is larger still. The sampled papers are not representative of the population of papers. The sample statistics are just that. No conclusion can be drawn about the level of consensus in the wider literature. The sampling strategy may have worked in favour or against the measured consensus on the hypothesis of anthropogenic climate change.

The conclusions of Cook et al. are unfounded. There is no doubt in my mind that the literature on climate change overwhelmingly supports the hypothesis that climate change is caused by humans. I have very little reason to doubt that the consensus is indeed correct. Cook et al., however, failed to demonstrate this. Instead, they gave further cause to those who believe that climate researchers are secretive (as data were held back) and incompetent (as the analysis is flawed).

Research benefits from the occasional stock-taking. Policy makers and other climate-research similarly benefit from surveys on the state of knowledge on climate change and its causes. Such reviews are better done by limiting the analysis to the relevant literature. The IPCC fulfills this role (Hegerl et al. 2007;Randall et al. 2007), and there are survey papers of both the model-based literature (Andrews et al. 2012), the palaeo-literature (Rohling et al. 2012) and the statistical literature (Annan and Hargreaves 2011). These surveys show that, indeed and without cooking the books, climate has changed and that greenhouse gases have played a substantial role in this.

Acknowledgements

I had useful discussions with Andrew K, Brandon Shollenberger, Dana Nuccitelli, Joshua Halpern, Lucia Liljegren, Shub Niggurath, and Willard.

References

Andrews, T., J.M.Gregory, M.J.Webb, and K.E.Taylor (2012), 'Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models', *Geophysical Research Letters*, **39**, (9).

Annan, J.D. and J.C.Hargreaves (2011), 'On the generation and interpretation of probabilistic estimates of climate sensitivity', *Climatic Change*, **104**, (3-4), pp. 423-436.

Cook, J., D.Nuccitelli, S.A.Green, M.Richardson, B.Winkler, R.Painting, R.Way, P.Jacobs, and A.Skuce (2013), 'Quantifying the consensus on anthropogenic global warming in the scientific literature', *Environmental Research Letters*, **8**, (024024).

Hegerl, G.C., F.W.Zwiers, P.Braconnot, N.P.Gillett, Y.Luo, J.A.Marengo Orsini, N.Nicholls, J.E.Penner, and P.A.Stott (2007), 'Understanding and attributing climate change', in *Climate Change 2007: The Physical Science Basis -- Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon et al. (eds.), Cambridge University Press, Cambridge, pp. 663-745.

Lyberg, L.E. and P.P.Biemer (2008), 'Quality assurance and quality control in surveys', in *International Handbook of Survey Methodology*, E.D. de Leeuws, J.J. Hox, and D. Dillman (eds.), Psychology Press, Abingdon.

Mohler, P.P., B.-E.Pennell, and F.Hubbard (2008), 'Survey documentation: Toward professional knowledge management in sample surveys', in *International Handbook of Survey Methodology*, E.D. de Leeuws, J.J. Hox, and D. Dillman (eds.), Psychology Press, Abingdon.

Randall, D.A., R.A.Wood, S.Bony, R.Colman, T.Fichefet, J.Fyfe, V.Kattsov, A.Pitman, J.Shukla, J.Srinivasan, R.J.Stouffer, A.Sumi, and K.E.Taylor (2007), 'Climate models and their evaluation', in *Climate Change 2007: The Physical Science Basis -- Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, S. Solomon et al. (eds.), Cambridge University Press, Cambridge, pp. 589-662.

Rohling, E.J., A.Sluijs, H.A.Dijkstra, P.Koehler, R.S.W.van de Wal, A.S.Von Der Heydt, D.J.Beerling, A.Berger, P.K.Bijl, M.Crucifix, R.Deconto, S.S.Drijfhout, A.Fedorov, G.L.Foster, A.Ganopolski, J.Hansen, B.Haenisch, H.Hooghiemstra, M.Huber, P.Huybers, R.Knutti, D.W.Lea, L.J.Lourens, D.Lunt, V.Masson-Demotte, M.Medina-Elizalde, B.Otto-Bliesner, M.Pagani, H.PÃ¤like, H.Renssen, D.L.Royer, M.Siddall, P.Valdes, J.C.Zachos, and R.E.Zeebe (2012), 'Making sense of palaeoclimate sensitivity', *Nature*, **491**, (7426), pp. 683-691.

Singer, E. (2008), 'Ethical issues in surveys', in *International Handbook of Survey Methodology*, E.D. de Leeuws, J.J. Hox, and D. Dillman (eds.), Psychology Press, Abingdon.

Table 1. One-sided deviations outside the 95% confidence interval for three indicators (standard deviation, first-order autocorrelation, skewness) for rolling windows of alternative widths (50, 100, 250 abstract).

	50-abstract		100-abstract		250-abstract	
	<2.5%	>97.5%	<2.5%	>97.5%	<2.5%	>97.5%
Standard deviation	2.8%	4.1%	3.1%	3.4%	4.3%	5.8%
Autocorrelation	1.6%	4.7%	1.4%	6.4%	0.0%	8.2%
Skewness	3.8%	2.2%	6.1%	3.3%	9.7%	4.5%

Table 2. Abstracts by subject and rating

Subject\rating[a]	1	2	3	4	5	6	7
Impacts	12	316	907	4528	8	5	4
Mitigation	20	418	1474	1471	1	2	0
Methods	28	161	391	1359	42	7	5
Palaeoclimate	4	27	138	612	3	1	0

[a] 1: explicit endorsement with quantification; 2: explicit endorsement without quantification; 3: implicit endorsement; 4: neutral; 5: implicit rejection; 6: explicit rejection without quantification; 7: explicit rejection with quantification

Table 3. Levels of endorsement[a]

	All	Explicit
All papers	98.0%	97.6%
Impacts + mitigation	99.4%	98.6%
Methods + palaeoclimate	92.8%	94.4%
Methods + palaeoclimate corrected[b]	89.9%	93.8%

[a] The level of endorsement is defined as the number of papers in columns 1, 2 and 3 in Table 1 over the number of papers in columns 1, 2, 3, 5, 6, and 7. Explicit endorsement omits columns 3 and 5.

[b] The correction is based on the assumption that methods and paleoclimate papers were misrated in the same proportion as impacts and mitigation papers.

Figure 1. Levels of endorsement in all papers, papers on impacts and mitigation, and papers on methods and palaeoclimate (left axis) and share of impacts and mitigation papers in total (right axis).

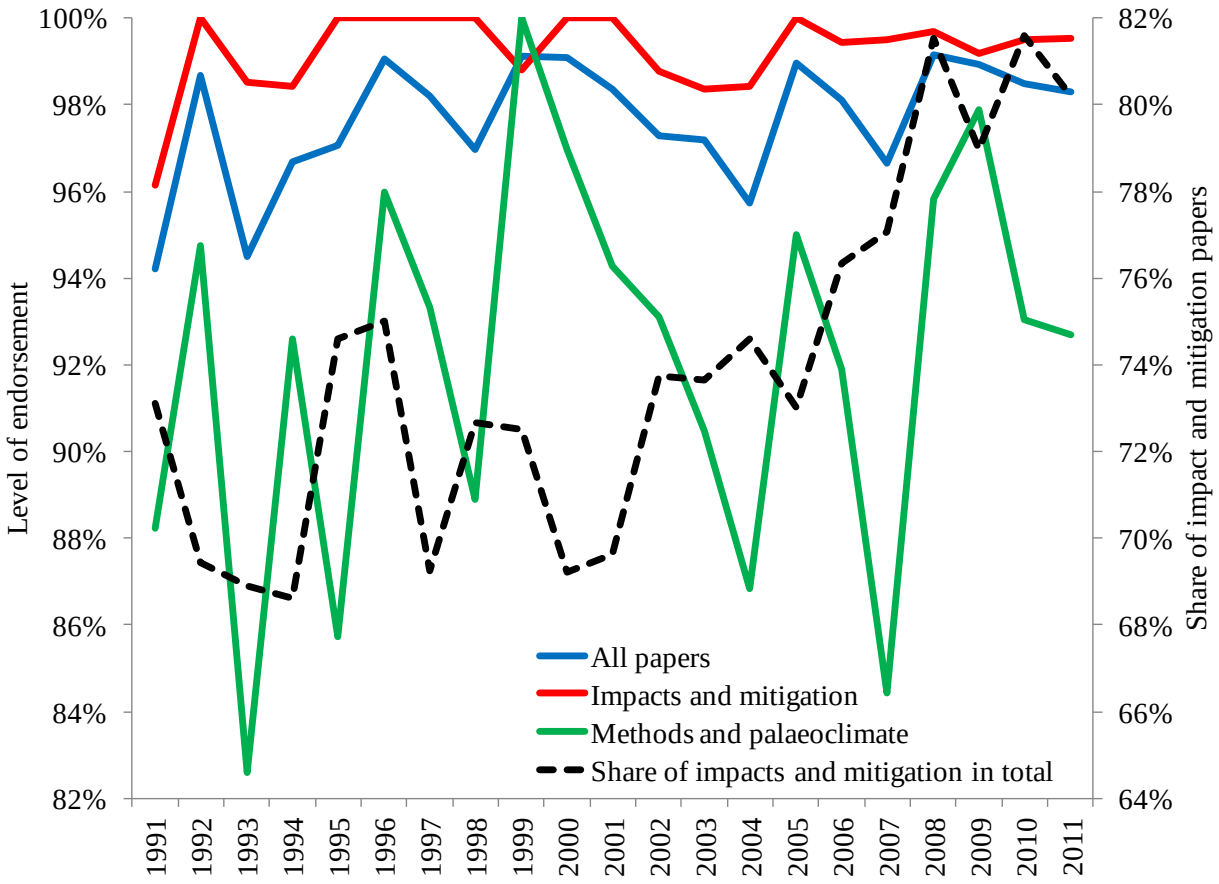


Figure 2. Relative and absolute deviations between the larger sample (“climate change”) and the smaller sample (“global climate change”) by discipline; positive numbers indicate oversampling.

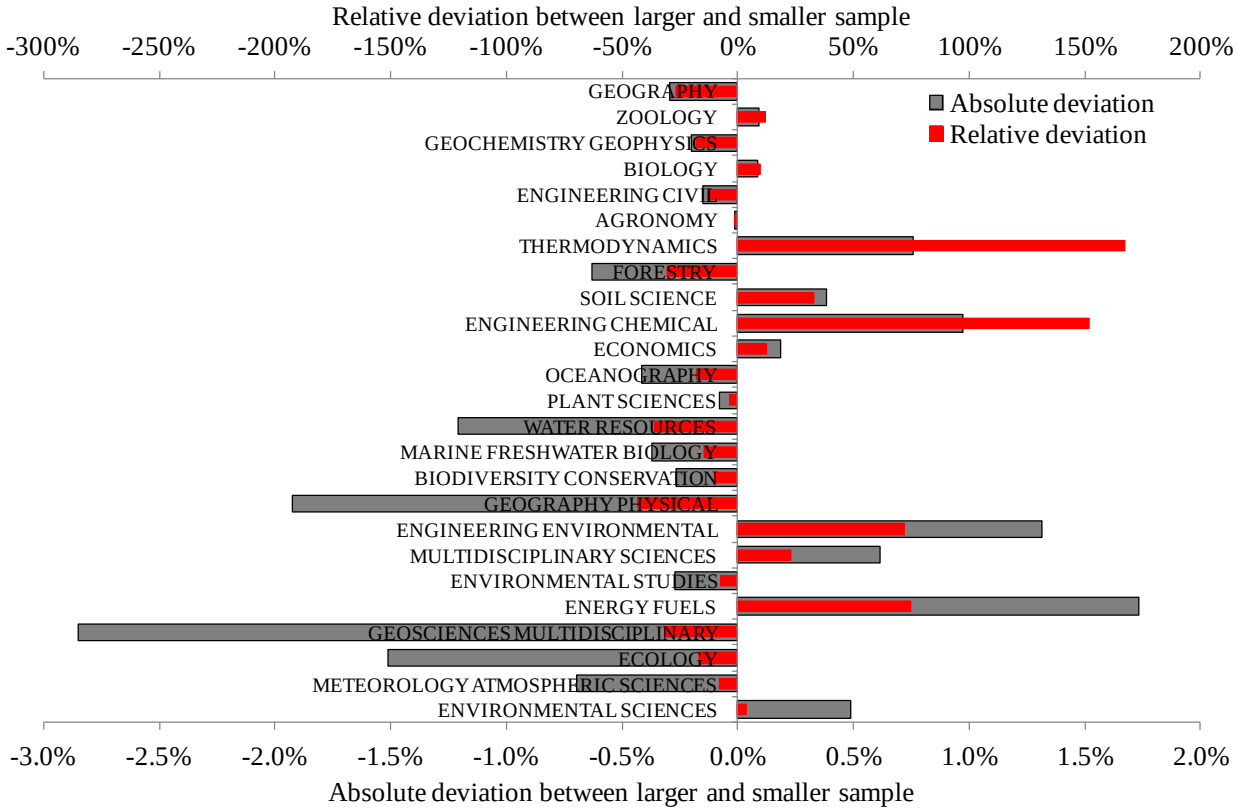


Figure 3. Number of papers by the 50 most prolific authors in the larger sample ("climate change"; red plus blue) and in the smaller sample ("global climate change", blue)

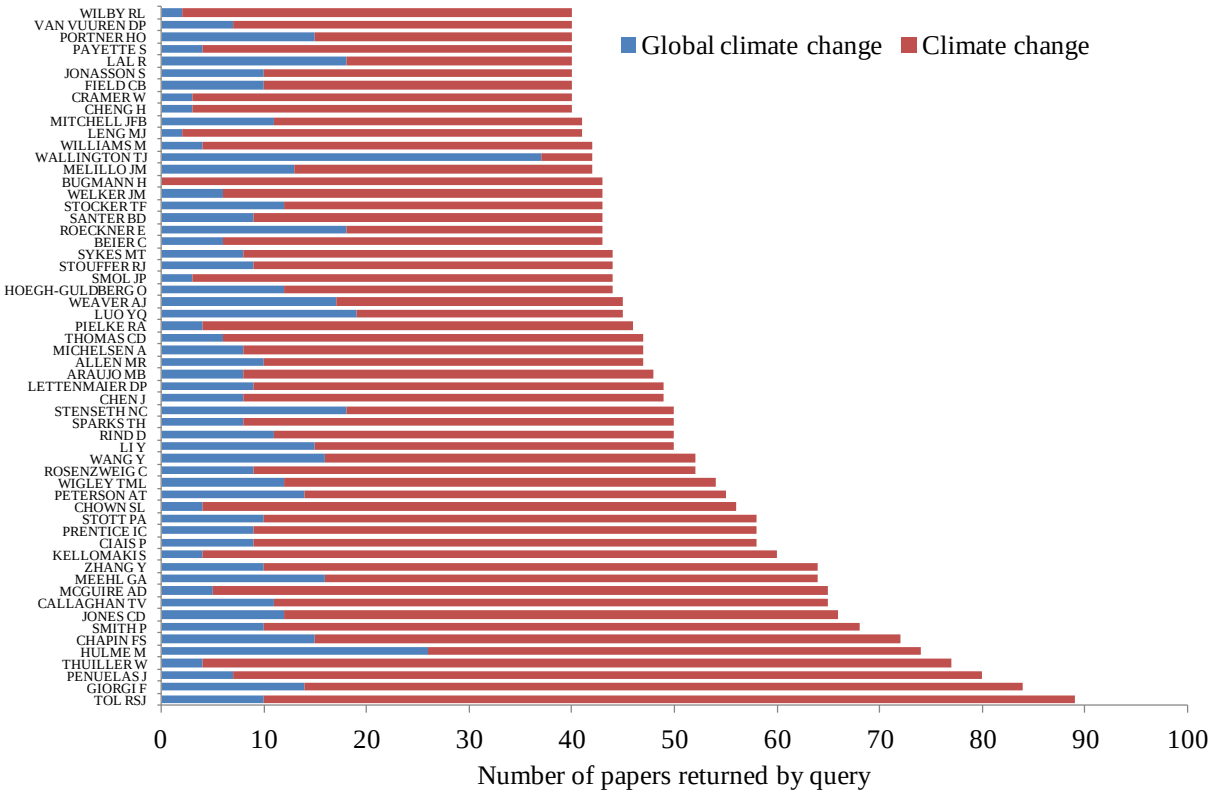


Figure 4. The fifty most-cited papers in the larger sample ("climate change"; red and blue); the papers also included in the smaller sample ("global climate change") are in blue.

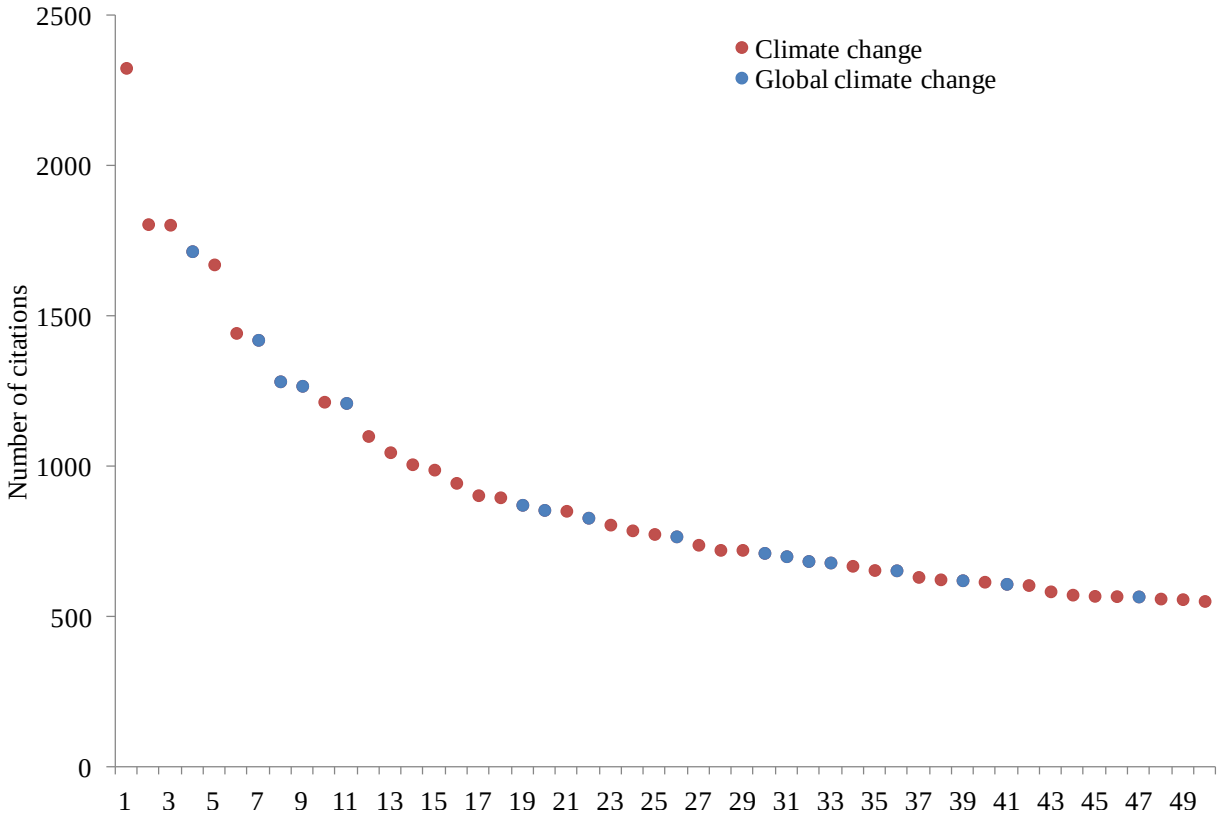
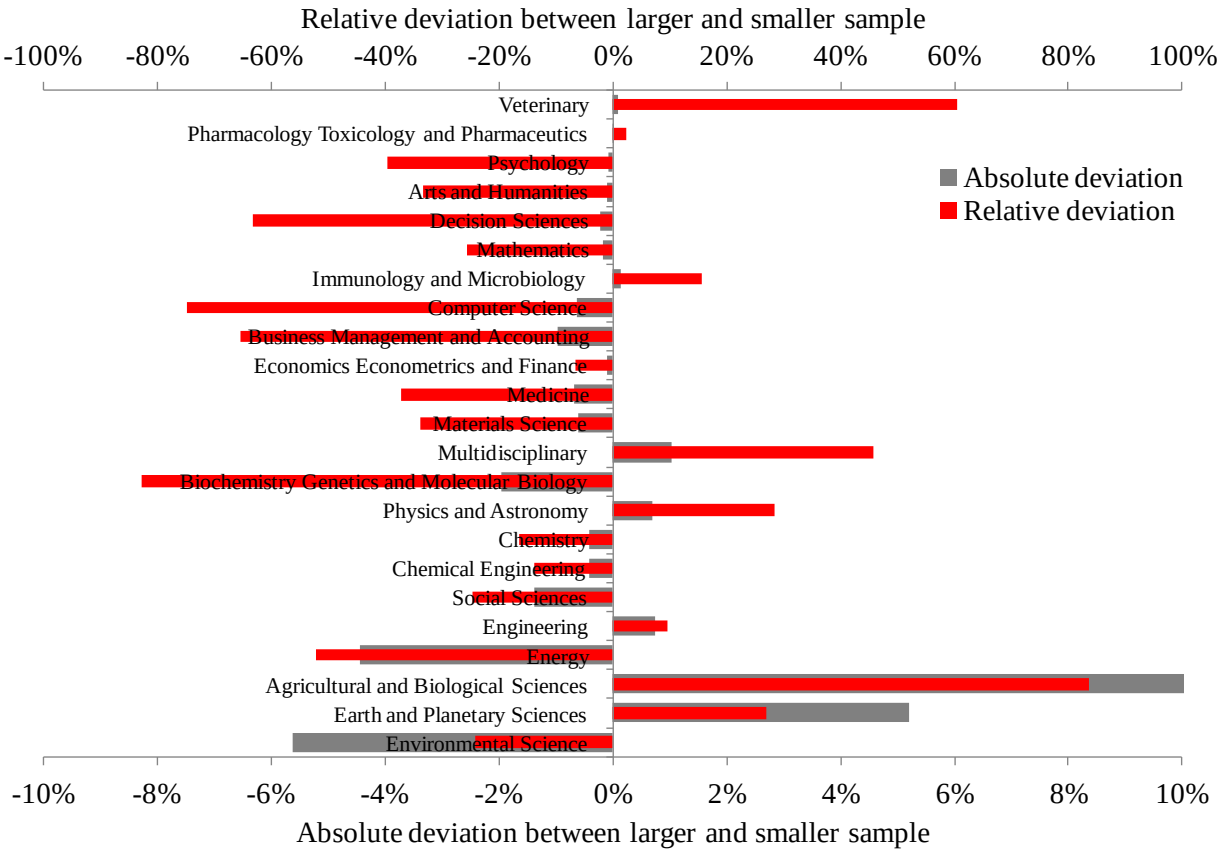


Figure 5. Relative and absolute deviations between the larger sample (Scopus) and the smaller sample (Web of Science) by discipline; positive numbers indicate oversampling.



SUPPLEMENTARY INFORMATION

Figure S1. Rolling standard deviation, 50-abstract window

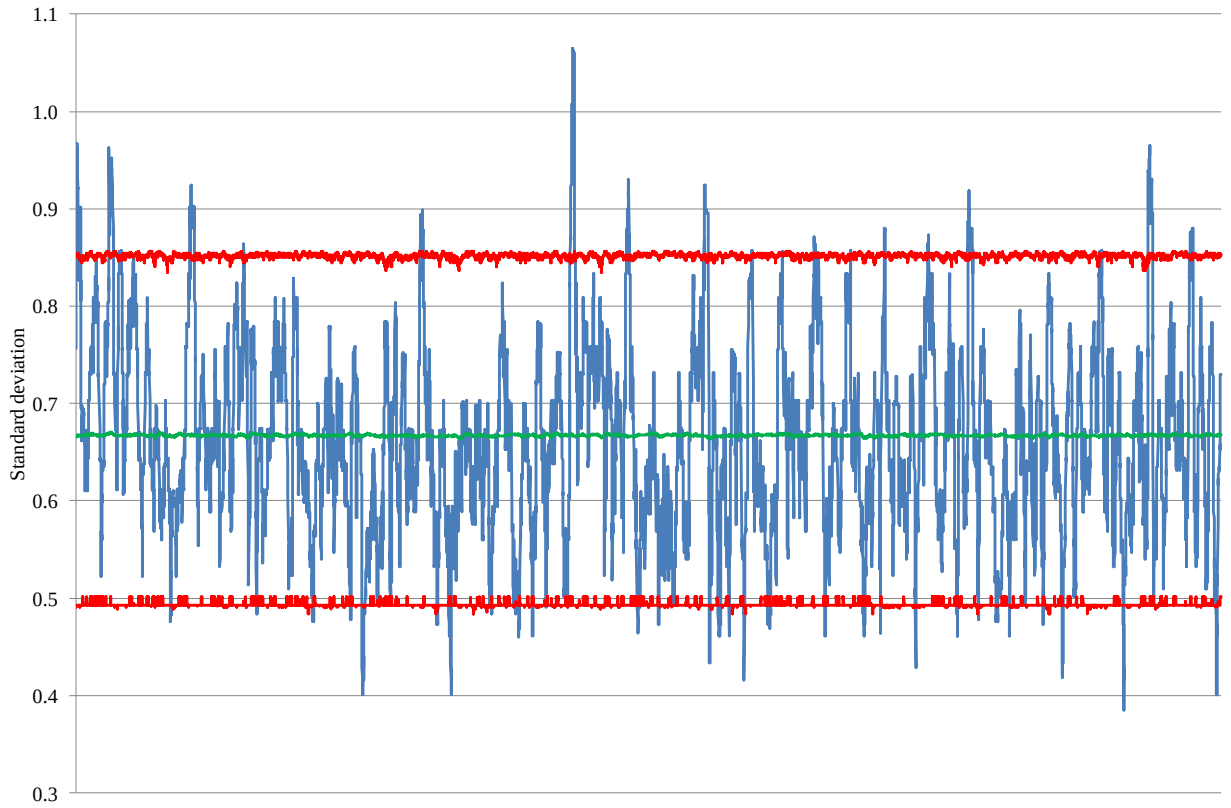


Figure S2. Rolling standard deviation, 100-abstract window

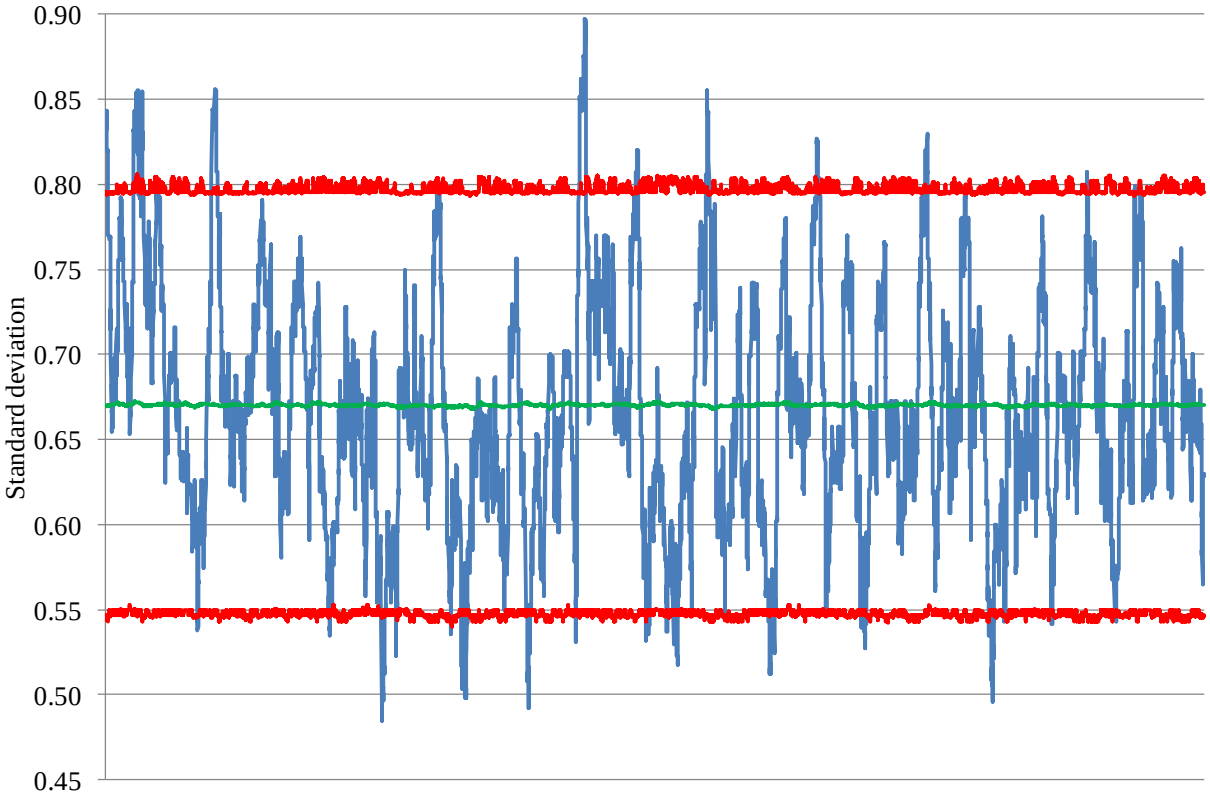


Figure S3. Rolling standard deviation, 500-abstract window

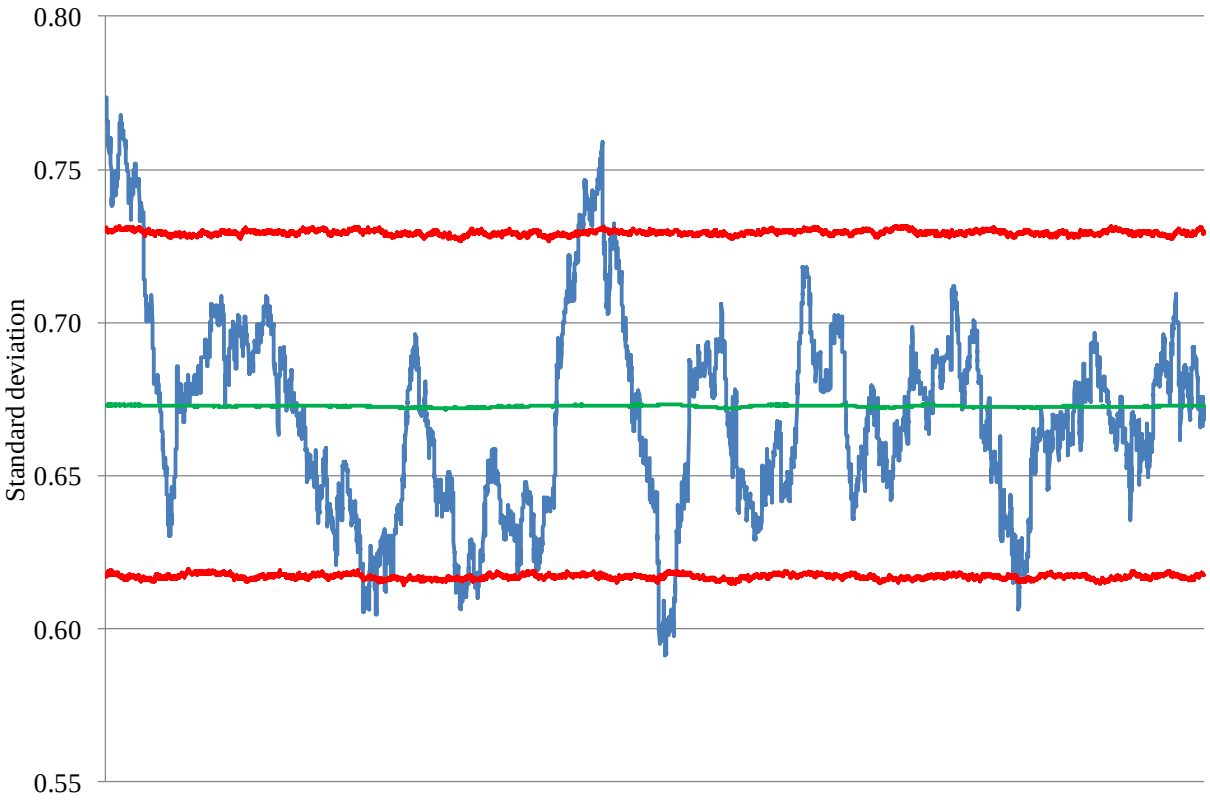


Figure S4. Rolling autocorrelation, 50-abstract window

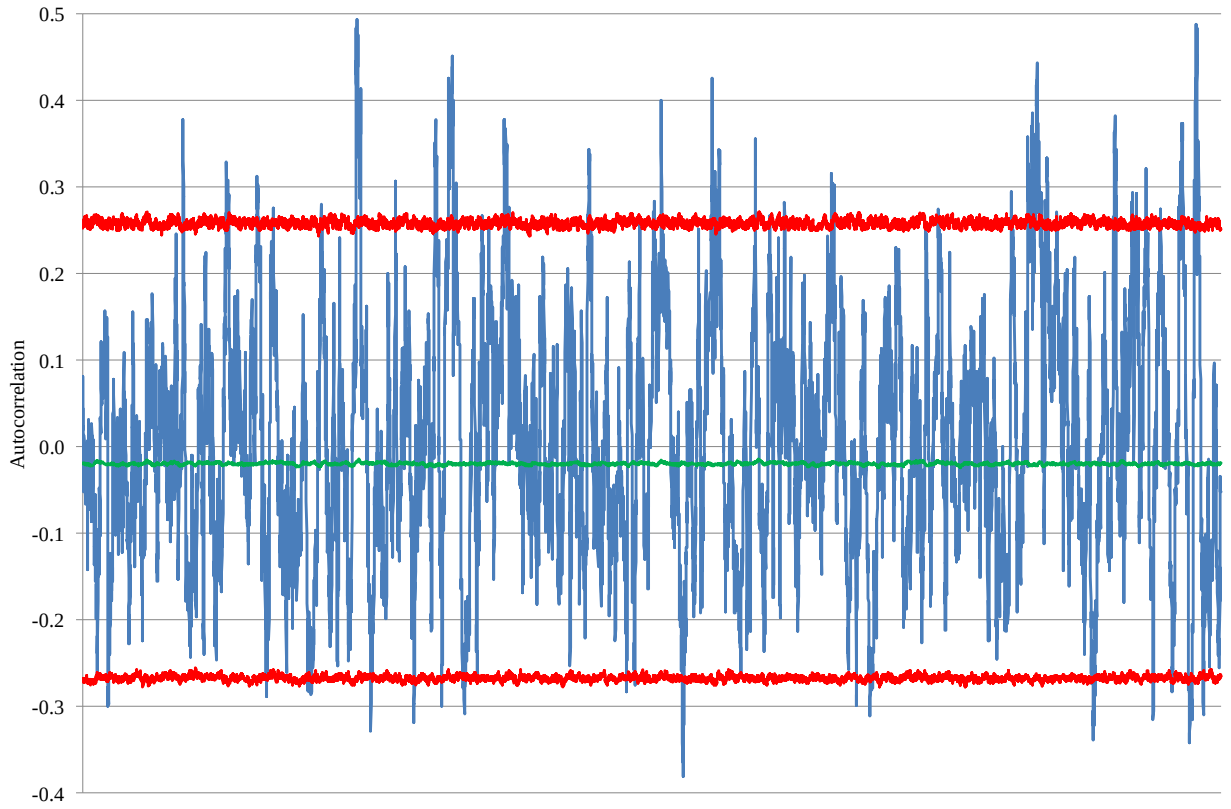


Figure S5. Rolling autocorrelation, 100-abstract window

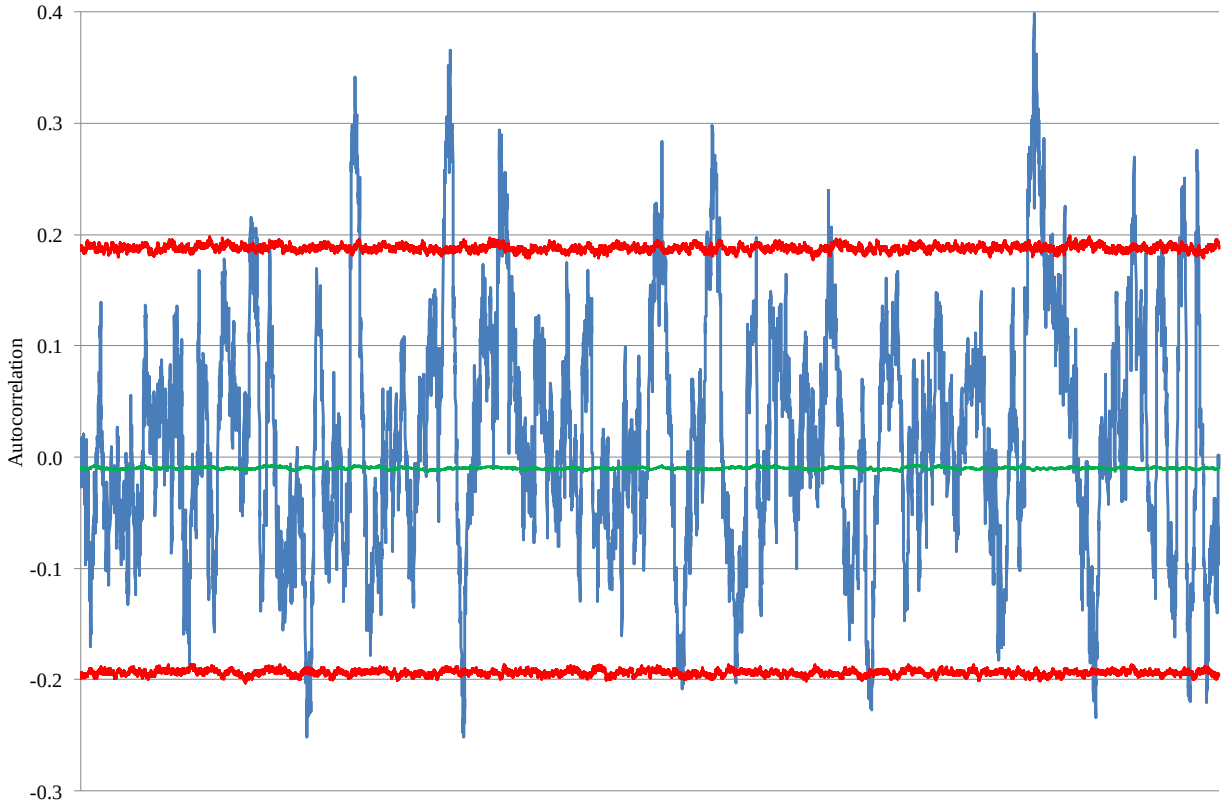


Figure S6. Rolling autocorrelation, 500-abstract window

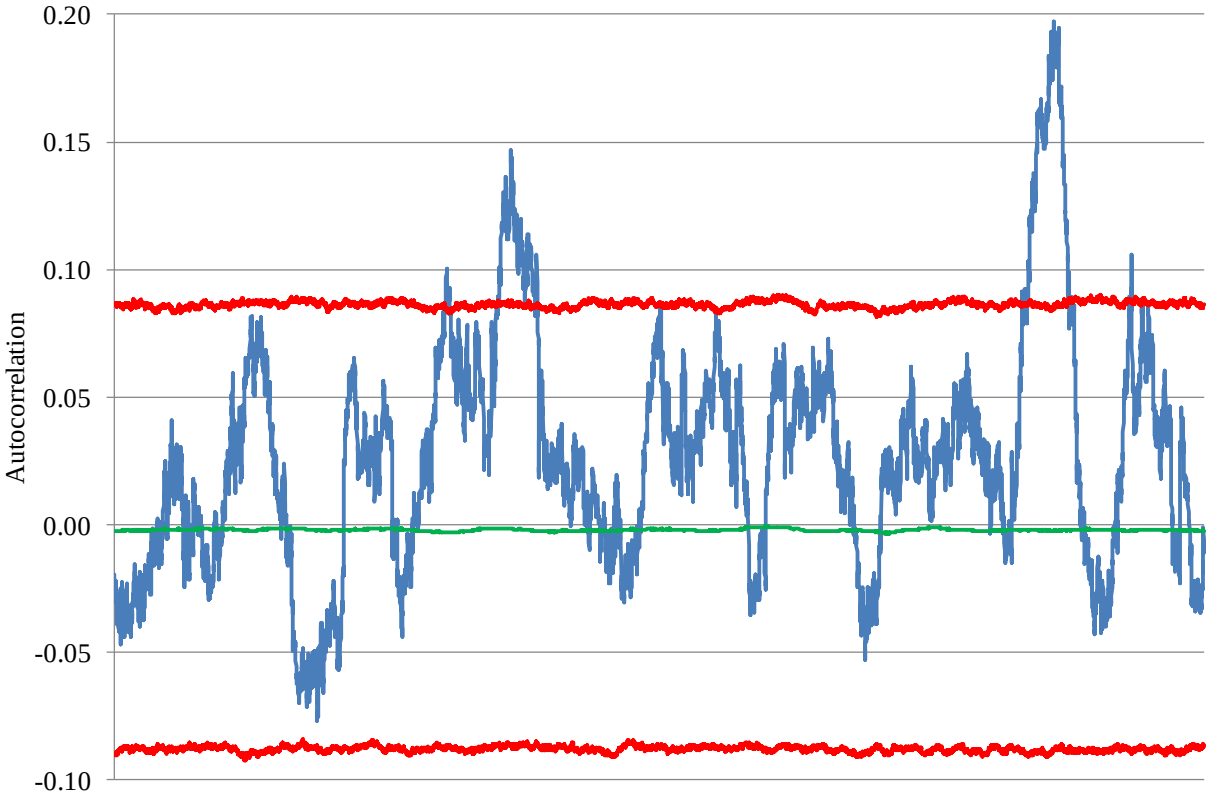


Figure S7. Rolling skewness, 50-abstact window

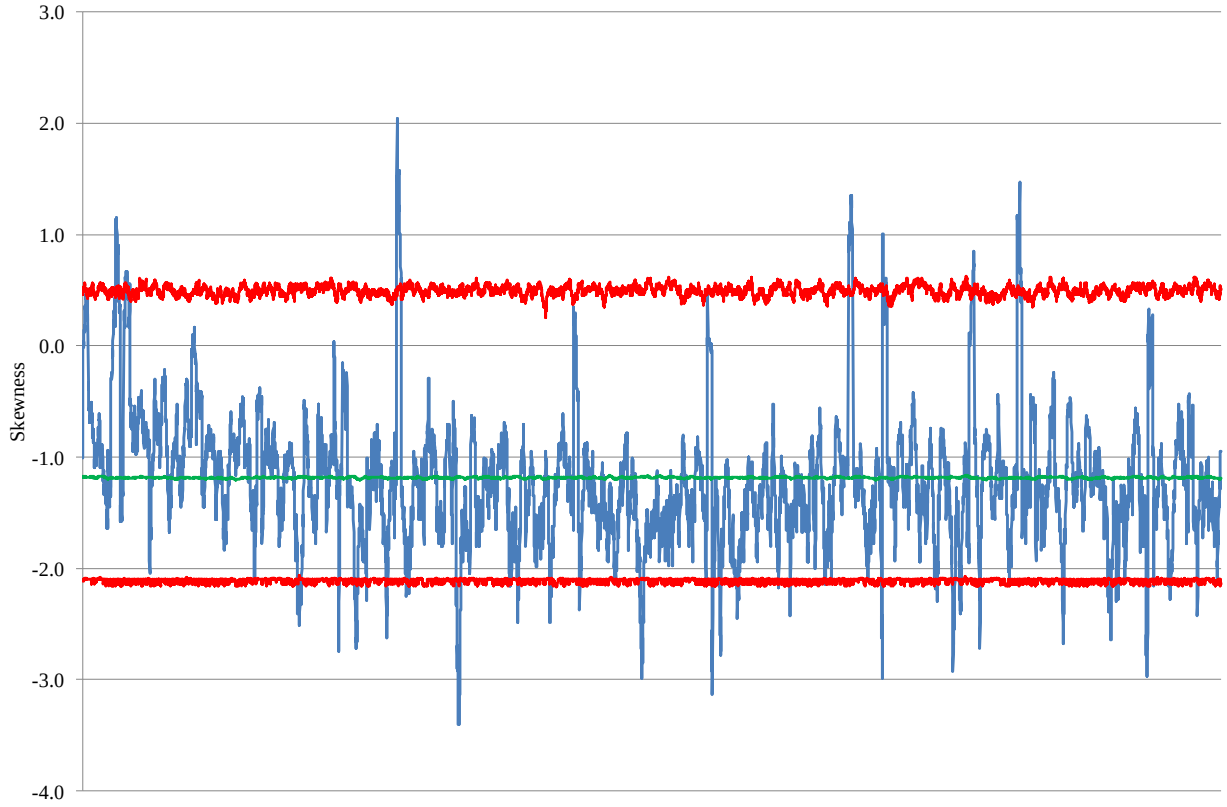


Figure S8. Rolling skewness, 100-abstract window

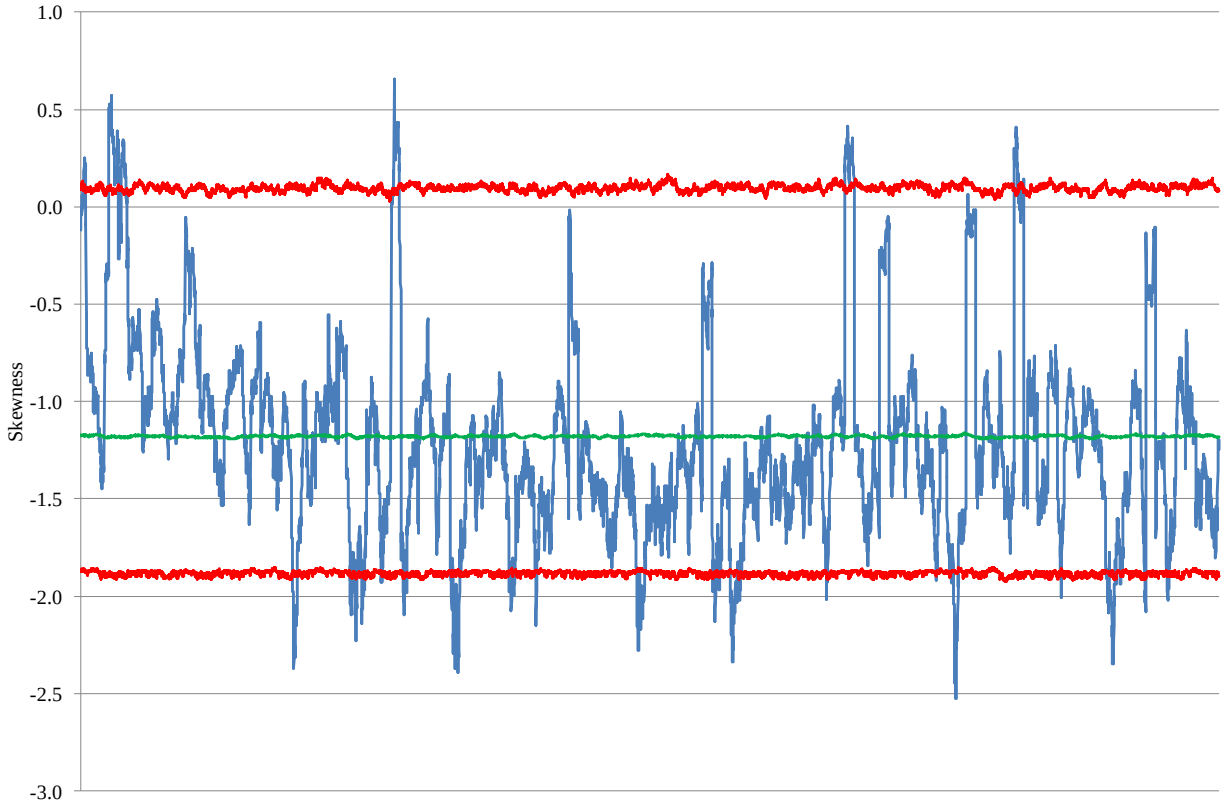


Figure S9. Rolling skewness, 500-abstract window

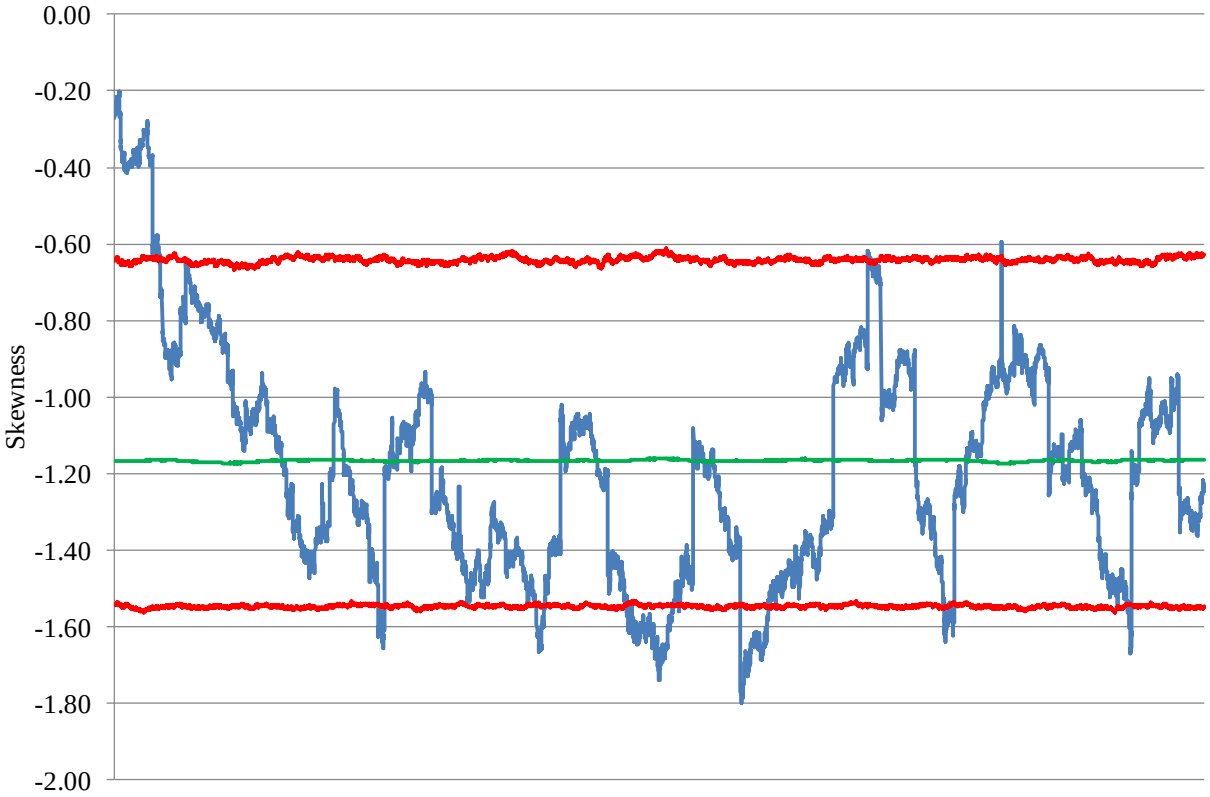


Figure S10. Autocorrelation function of the ratings in the reported, part-alphabetical order and in alphabetical order; the dashed lines denote the 95% bootstrapped confidence interval.