REPLY • OPEN ACCESS

Reply to Comment on 'Are physicists afraid of mathematics?'

To cite this article: Jonathan E Kollmer *et al* 2016 *New J. Phys.* **18** 118004

View the article online for updates and enhancements.

New Journal of Physics

The open access journal at the forefront of physics

CrossMark

REPLY

Reply to Comment on 'Are physicists afraid of mathematics?'

Jonathan E Kollmer[1,2], Thorsten Pöschel[1,2,3,4] and Jason A C Gallas[1,2,3,4]

1. Institute for Multiscale Simulations, Friedrich-Alexander-Universität, D-91052 Erlangen, Germany
2. Max-Planck-Institut für Physik komplexer Systeme, D-01187 Dresden, Germany
3. Departamento de Física, Universidade Federal da Paraíba, 58051-970 João Pessoa, Brazil
4. Instituto de Altos Estudos da Paraíba, Rua Silvino Lopes 419-2502, 58039-190 João Pessoa, Brazil

Abstract

Based on citation data of biologists and physicists, we reiterate that trends in statistical indicators are not reliable to unambiguously blame mathematics for the existence or lack of paper citations. We further clarify that, contrary to claims in the Comment (Higginson and Fawcett 2016 *New J. Phys.* **18** 118003), a clear statistical correlation between the number of equations and the citation success is not possible because the data is too noisy and not reliable for identifying trends unambiguously. Concerning their conclusions, we stress the well-know fact in statistics that even if correlation could be found, it by no means implies causality. Accordingly, to discuss ways of increasing citation rates by suppressing or hiding equations in appendices cannot be justified with statistics, even less so when based on small sets of very noisy data.

In a recent paper [1], we considered the correlation between citation rates and several formal features of scientific papers in the field of physics. Among these features was the number of display equations contained in the text. The database used in our analysis consisted of two volumes of *Physical Review Letters*, comprising about 2,000 papers. In our analysis, we could not identify a clear correlation between the number of equations and the citation success because, essentially, the data is so noisy that it is not reliable for identifying trends *unambiguously*.
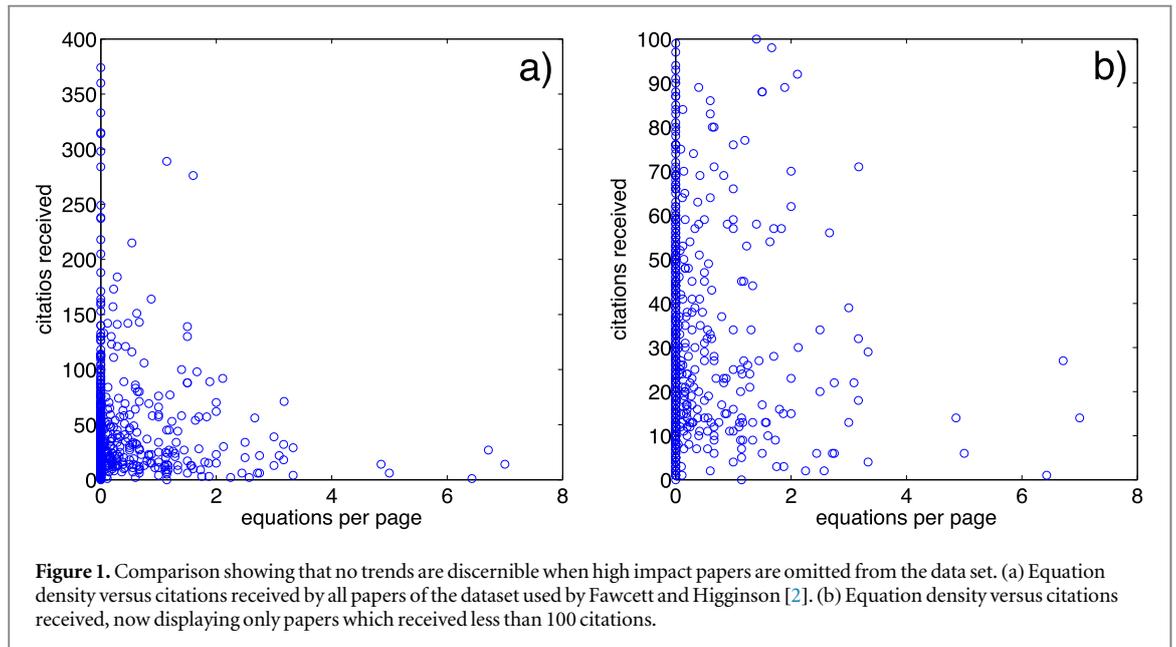
Our conclusion is at odds with earlier findings by Fawcett and Higginson [2] who analyzed papers in biology and claimed to observe a negative correlation, a fact that led them to the conclusion that 'heavy use of equations impedes communication among biologists' and to speculations about how one could 'enhance communication'.

Our original paper [1] also re-analyzed the data set used by Fawcett and Higginson [2] (provided by them as supplementary material) and could not confirm their findings. In particular, their result was found to depend strongly on the details of the statistical analysis, a point also raised by Gibbons [3]. In this reanalysis, we found conclusions drawn by Fawcett and Higginson to be strongly dependent on their subdivision of papers into theoretical and non-theoretical work, which they based on matching the keyword 'model' to the title of a paper. We also pointed out that the data is quite sparse and often dominated by single high impact papers, especially in the light that the already sparse data is then further subdivided in citations coming from either 'theoretical' or 'non-theoretical' papers (which assumes that papers are either theoretical or not, that is, not accounting for papers containing theoretical as well as non-theoretical content). On the other hand, looking at the individual papers one clearly finds examples of highly successful equation rich papers such as the in fact two of the most cited papers that in the dataset analyzed by HF. One of them contains 1.9 equations per page and was cited 300 times, with 118 citations coming from 'non-theoretical' papers, while another features 2.7 equations per page, and received 291 total citations, with a majority of 182 arising from 'non-theoretical' papers.

Now, in a Comment to our paper, Higginson and Fawcett [4] express several opinions focusing on two points of the aforementioned issues: (a) that our 'analysis is flawed and the claims are unsupported', and (b) draw strong conclusions concerning methods to increase citation success which they believe are erroneous. In this Reply we address both points in some detail.

Table 1. When the 'Negative Binomial model' used by [2] is applied to the data set that they analyzed in [2] excluding papers cited more than 100 times, a correlation between equation density and citation rate cannot be substantiated. As usual, asterisks indicate significance of the numbers [5].
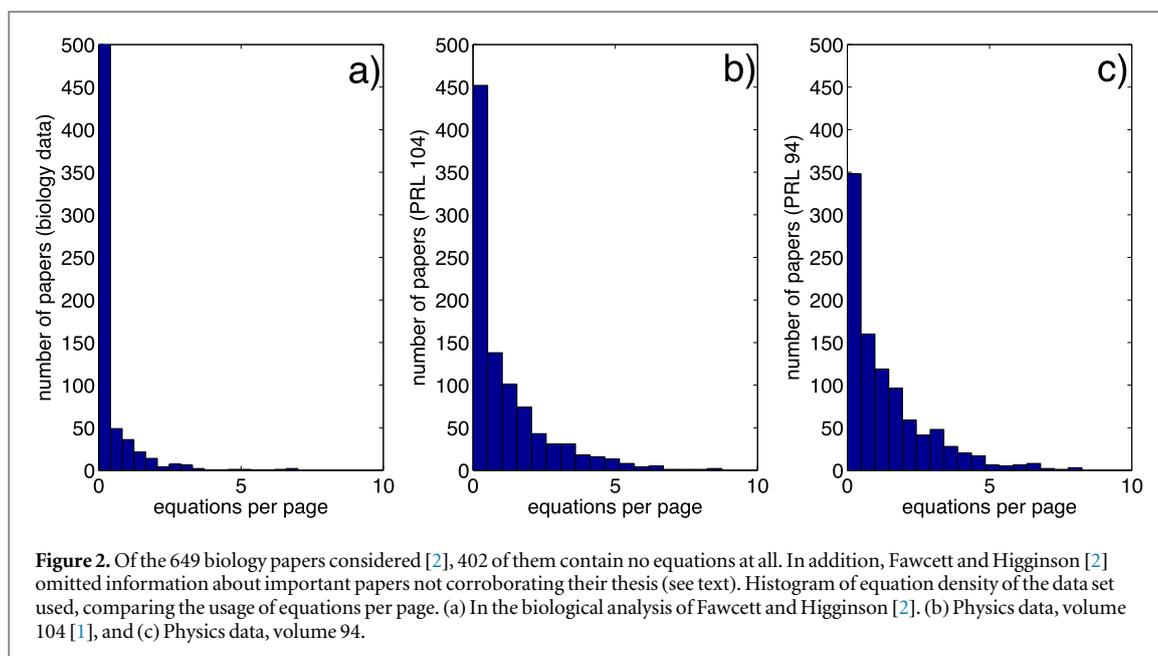
	Estimate	Std Error	z	P
(Intercept)	3.195	0.144	22.143	<2e-16***
equation density	−0.101	0.073	−1.370	0.171
Pages	0.031	0.011	2.819	0.00481 **
Journal Evolution	−0.068	0.097	−0.703	0.48237
Journal Proc B	−0.025	0.101	−0.252	0.80132
equation density × pages	0.006	0.009	0.696	0.48620

Figure 1. Comparison showing that no trends are discernible when high impact papers are omitted from the data set. (a) Equation density versus citations received by all papers of the dataset used by Fawcett and Higginson [2]. (b) Equation density versus citations received, now displaying only papers which received less than 100 citations.

Statistical analysis: First, in contrast to what Higginson and Fawcett write, our paper [1] pointed out that we did not find evidence for a correlation between citation frequency and density of equation. This is quite different from what Higginson and Fawcett claim, stating erroneously that we denied the existence of correlations, which is not the case. Instead, we showed that the results depend very sensitively on the details of the analysis. For, modifying the method only slightly, the end result may change qualitatively. For instance, in the present Reply our data used in [1] was re-analyzed. In order obtain strong significance of the correlation, their table 1—see column (B) of Higginson and Fawcett [4]—had to exclude those papers which earned more than 100 citations from the analysis. If we apply the *same* standard to their data used in [2], the correlation reported there vanishes (*P*-value of 0.171, see our table 1), that is, a correlation cannot be substantiated. In particular, as may be seen from figure 1, no trends are discernible from the data set when high impact papers are omitted. From all this, we conclude that the statistical analysis is dominated by a small number of heavily cited papers.

Second, it is important to realize that only a small fraction of the papers analyzed by Fawcett and Higginson [2] contains a substantial amount of displayed equations. Namely, of the **649** papers analyzed by them, **411** papers contain no equations at all. Furthermore, only 85 papers contain one or more equations per page, as illustrated here in figure 2. That is, the effective data set of relevant papers is significantly smaller than claimed in [2]. Concerning Fawcett and Higginson [2] not mentioning data strongly contradicting their views, they now explain [4]: 'Statistically, nothing can be inferred from two examples of heavily cited, equation dense outliers selected from a sample of 649 articles.' In fact, as it is evident from our figure 2, that the three datasets under discussion consist basically of what they are calling 'outliers'.

Third, for their analysis of the data about physics, Higginson and Fawcett combined both volumes of PRL to form a single data set. We find this at least questionable, since both volumes were published with a temporal separation of 10 years, the publication and citation habits might have changed significantly. Considering both volumes as a single data set would require prior statistical investigation to justify it. Moreover, in their Comment, Higginson and Fawcett once again introduce an *interaction term* between the journal volume and the

Figure 2. Of the 649 biology papers considered [2], 402 of them contain no equations at all. In addition, Fawcett and Higginson [2] omitted information about important papers not corroborating their thesis (see text). Histogram of equation density of the data set used, comparing the usage of equations per page. (a) In the biological analysis of Fawcett and Higginson [2]. (b) Physics data, volume 104 [1], and (c) Physics data, volume 94.

density of equations (see also the Comment by Gibbons [3] on the use of interaction terms in [2]). We point out that the correlation is much less significant when analyzing the volumes separately.

Correlation versus causality: The second part of the Comment concerns *interpretation of the data*. Here, Higginson and Fawcett draw conclusions which would require a *causal relation* between citation frequency and density of equation. However, what was investigated in [1] is a *correlation* and by no means a causality relation. That correlation does not imply causation is a well-known fact [6, 7]. Concerning this topic, see also the Comment by Fernandes [8] regarding the paper of Fawcett and Higginson [2]. Therefore, we do not understand the purpose of Higginson and Fawcett stating things like '*the 45 articles in data* [1] *that are moderately well cited (50–100 citations) and equation dense* (⩾2 *equations per page) would have (all else being equal) attracted an additional 476 citations (17% of their total) if the authors had halved the density of equations in the main text.*' To us, this is a prediction that seems hard to verify with the data given since it is statistically impossible to imply causality by correlation.

A known textbook example is the fact that cigarette smokers are murdered more frequently than non-smokers [9]. However, this is not reason enough to conclude that the total amount of murders would decrease if everybody would stop smoking. In the same way, even if it would be possible to find a negative correlation between citation frequency and density of equations (the fact we dispute), this would not allow one to conclude that the total number of citations would increase if the authors would reduce the density of equations. Again, this is a well-know fact in statistics [6, 7]: correlation does not imply causality, and we believe that the conclusions drawn by Higginson and Fawcett are based on violation of this rule.

Consequently, if one could find the correlation claimed originally [2] and repeated in the present Comment [4], several reasons could be responsible for them. For example, different areas in any discipline (physics, biology) require different level of mathematical description and it seems quite natural that scientists working in a certain area cite preferred papers published in the same area. Since for several reasons the theoretical research areas are traditionally less populated, it appears quite naturally that there are differences in the citation rates which, however, cannot be attributed to 'impeded communication' as claimed in [2] and the current Comment [4]. Whether or not this plausible explanation is correct cannot be evidenced by pure statistical analysis but would need sociological and more elaborated research.

In summary, we believe the main thrust of our work holds true, that is, a correlation between the citation rate and the number of display equation cannot be evidenced by a statistical analysis based on the available data, dominated too much by noise. We find very disquieting Higginson and Fawcett to be 'using the most appropriate statistical analysis' [4] while working with rather small samples of data which contain papers cited in a way obviously contradicting their thesis. We find the statistical analysis of citation data to be dominated by a small number of heavily cited papers, whose success is more likely attributable to their scientific content than to the number of equations used or to other secondary points. Thus, our work points out that if you want to draw these strong conclusions a much larger set of data should be analyzed, reducing the need to rely on details of very specialized *ad-hoc* statistical analysis for sparse data. The fact that millions of scientific articles are available online and analysis is automated shows that there is no need to restrain oneself to these small sample sizes.

Finally, we do not understand the connection of our paper with the purpose of their statement concerning the need for presenting papers in an accessible manner by reiterating that 'all scientists aiming to communicate theory in the most effective way should take this issue seriously, rather than claiming it does not exist'. Our paper does not address such claim.

Acknowledgments

References

[1] Kollmer J E, Pöschel T and Gallas J A C 2015 Are physicists afraid of mathematics? *New J. Phys.* **17** 013036
[2] Fawcett T W and Higginson A D 2012 Heavy use of equations impedes communication among biologists *Proc. Nat Acad. Sci. USA* **109** 11735
[3] Gibbons J 2012 Do not throw equations out with the theory bathwater *Proc. Natl Acad. Sci. USA* **109** E3054
[4] Higginson A D and Fawcett T W 2016 Comment on 'Are physicists afraid of mathematics' *New J. Phys.* **18** 118003
[5] Hoem J M 2008 The reporting of statistical significance in scientific journals *Demographic Res.* **18** 437–42
[6] Huff D 1954 *How to Lie with Statistics* (New York: W.W. Norton)
[7] Steele J M 2005 Darrell huff and fifty years of How to lie with statistics *Statistical Sciences* **20** 205–9
[8] Fernandes A D 2012 No evidence that equations cause impeded communication among biologists *Proc. Natl Acad. Sci. USA* **109** E3057
[9] Ghanbari S A 2002 *Einführung in die Statistik für Sozial- und Erziehungswissenschaftler* (Berlin: Springer) p 99