The Shape of a PhD: A Data-Driven Visual Guide of One Student's Postgraduate Experience

Author: George Lewis

Affiliation: Dept. of Materials Science & Metallurgy, University of Cambridge

Date: April 2023

People love to quantify stuff. We measure and compare in the hope that it will help us to understand and improve. This is especially true in science, where we are constantly trying to explain the world around us through making measurements. A typical PhD thesis in the sciences contains an extensive set of figures (my research group has averaged ~100 figures per thesis in recent years), and these figures aim to visualise measurements in such a way that complex ideas can be easily shared. This was a part of my PhD that I particularly enjoyed; I often found hours would fly past as I adjusted every last detail of a figure before deciding it sufficiently did justice to the measurements displayed within it.

In this article, I wish to use my measurement-making, data-displaying obsession to look back over my PhD and try to see what insights can be gathered. I start with a more qualitative overview of the major milestone dates during the PhD, and 'conventional' measures of research such as citation count, before moving onto less conventional metrics such as files created, emails sent, and the frequency of diary entries. It is my hope that this will be of some interest to other PhD students, and may help prospective students to get a flavour of what postgraduate life is like.

Milestone dates

A PhD typically allows students a great deal of flexibility in how to structure their time, though from **Figure *1*** we perhaps rather see that the PhD *forces* students to structure their own time. Looking at the 'Official' commitments of the PhD student, there is simply a check-up/review at the end of the first year, and then thesis submission and viva at the very end. Of course, supervisors and research groups can impose additional structure; but it many cases this is minimal, and so it really is down to the individual to structure their time.

Looking at the various important dates throughout my PhD, I realise that having small jobs on the side actually provided me with some of the structure that I was looking for, even though at the time I was worried they may be detracting my focus from research. For example, comparing the timings in **Figure *1*** I now realise that knowing I had a 3-month internship coming up was one of the things that spurred me on to finish writing my first paper – without this imposed deadline I would likely have taken much longer to finish off that chapter of work. Similarly, being engaged in various undergraduate teaching roles kept me in sync with the ebb and flow of university term-time and provided me with natural prompts to think about my plan for the upcoming term, or what writing I could get done in the comparatively peaceful undergraduate holidays. I see now that carving out pieces of my time to teach and to work in local companies actually left me with neat, structured stretches of time between that helped me to focus my research efforts.

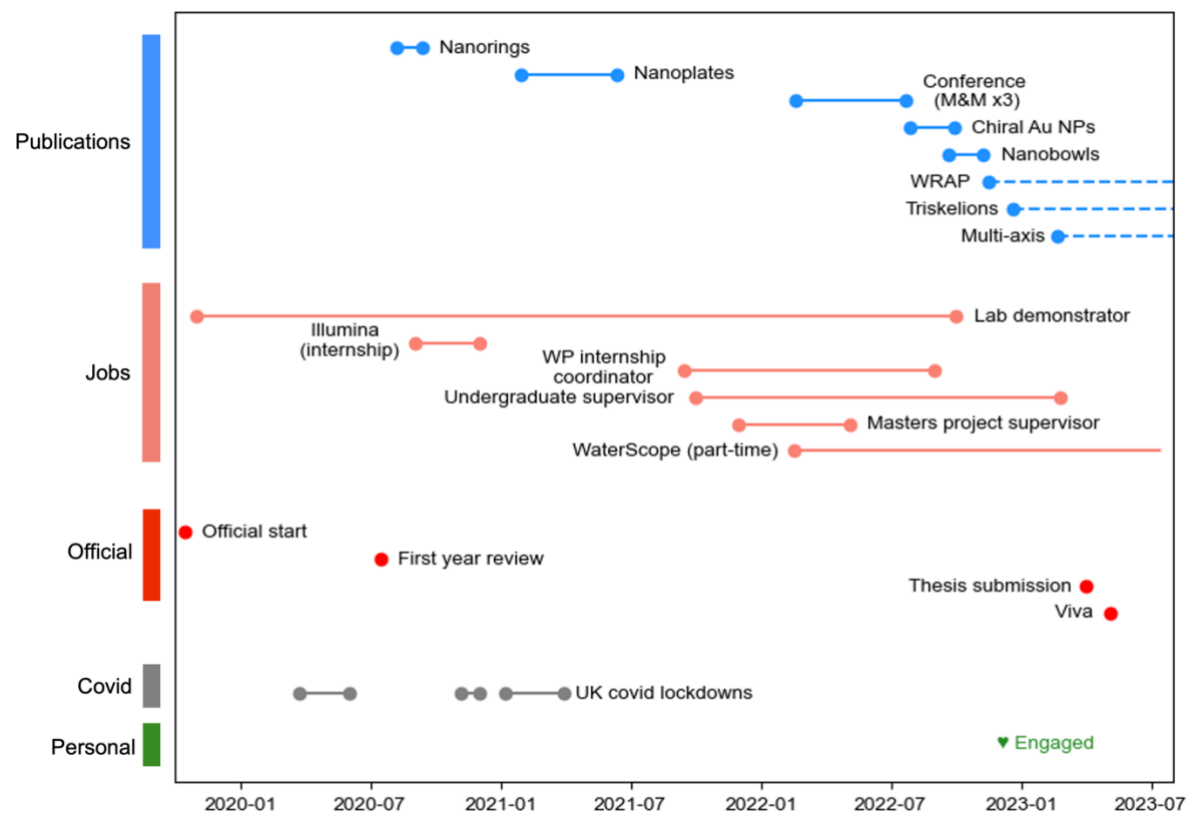

Figure 1 | PhD Milestones. Chart shows the duration/timings of major milestones throughout the PhD. Note that for publications, the duration indicates time between submission and acceptance, dashed lines indicate publications currently under review.

Another striking feature that **Figure *1*** shows is the pile-up of paper submissions in the final year of the PhD. You will often hear people tell you something along the lines of "*All the real work gets done at the end of your PhD*", and this is likely what they're talking about. This however is not the full picture, and is a topic I will return to later in the article.

Finally, whilst being hard to quantify or to track with milestones, PhD's are immensely fun. Long hours spent with colleagues allow friendships to form, flexible working patterns give you freedom to explore interests outside work, and many discussions at home with your partner such as "*Do you think this is a good colour scheme for my magnetic field lines?"* can only serve to strengthen a relationship over the course of a PhD.

Conventional research metrics

I of course am not the only one attempting to quantify research activity – in academia it is hard to escape discussions about number of publications, citation count, h-index and all the other factors that are used to keep track of academic output. Since this is the ecosystem of the academic world, it is naturally the easiest for us to keep track of. Through some simple web-scraping of ResearchGate's profile statistics, I was able to extract a monthly count of citations and 'reads' of my articles as shown in **Figure 2**.

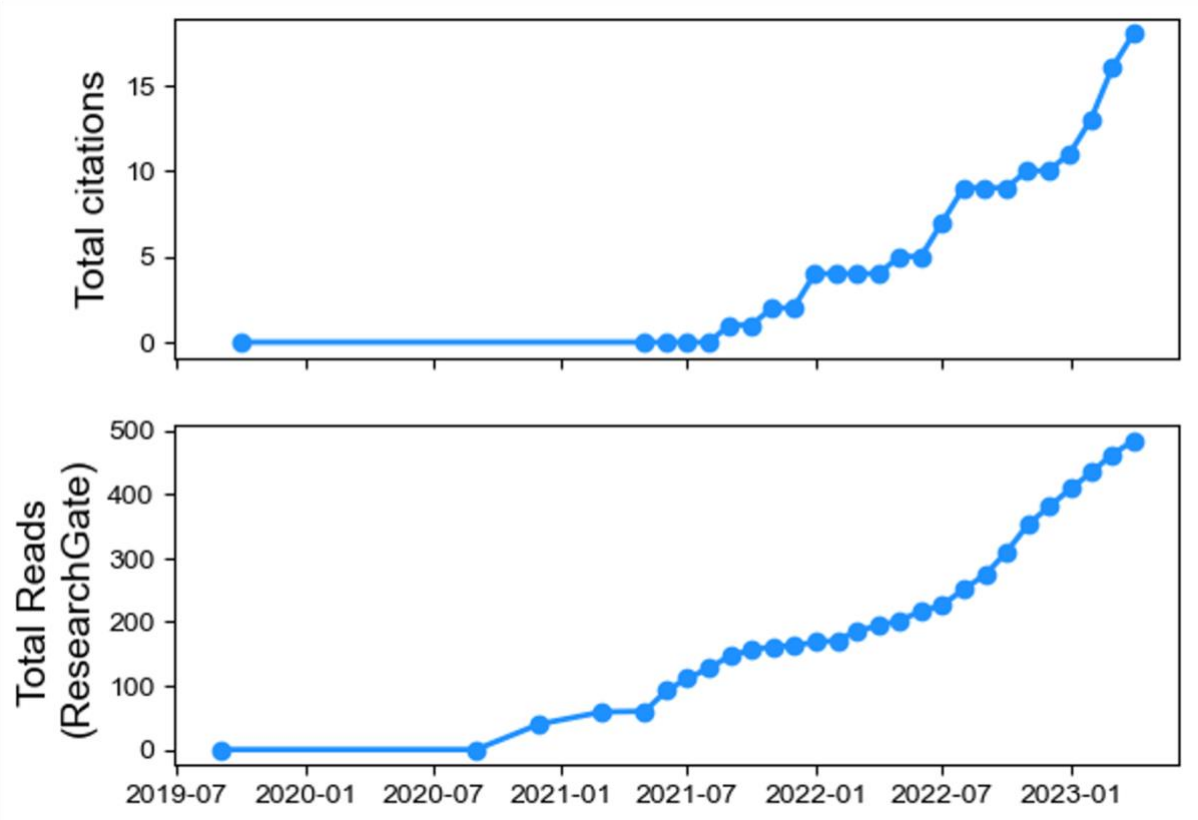

Figure 2 | Conventional metrics. (Top) Cumulative number of citations and (Bottom) number of article reads. Both datasets are extracted from ResearchGate profile statistics.

As anticipated, we see that both citation count and article reads only increase significantly in the final year of the PhD – so is it really true that only the end of your PhD is productive? Isn't that what we see in **Figure 2**?

The problem is that PhD students are at the beginning of their research career, and these conventional metrics are not designed for beginners. Mannella and Rossi found that the h-index only becomes a reliable metric 10 years after your first publication[1] – and even then, it remains strongly correlated with a researcher's academic age.

However, despite the clear inadequacy of these metrics in measuring beginner's output, it is hard as a PhD student not to measure your progress by these values. This can naturally lead to an uneasy feeling that you are not making enough progress, not putting in enough effort, or the well-known feeling of 'imposter syndrome'. Given these limitations, I thought I would try tracking my PhD progress using some different metrics, and in the next section, we'll see how these begin to fill in the parts of the picture that are missing so far.

Unconventional research metrics

Our interactions with computers leave behind a huge wealth of data. In this section I extract some of this data to see what it has to say about the way I carried out my PhD. Specifically, I start by looking at three metrics in **Figure 3**:

- First, the number of code commits made to PhD-related Github repositories (plotted cumulatively). Much of my research involved simulations and algorithm development, and so the frequency of code uploads could reasonably be interpreted as a measure of productivity.

- Second, the number of files created and stored on OneDrive (plotted cumulatively). I used OneDrive to back-up all my files throughout the PhD, so this metric keeps track of things like presentations made, reports written, figures made, and data stored; so could act as a second measure of productivity.
- Third, the number of emails sent and received (plotted month-to-month). A large part of this may be admin related, but it is also potentially a proxy measure for collaboration levels with other researchers, and acts as a third measure of productivity.

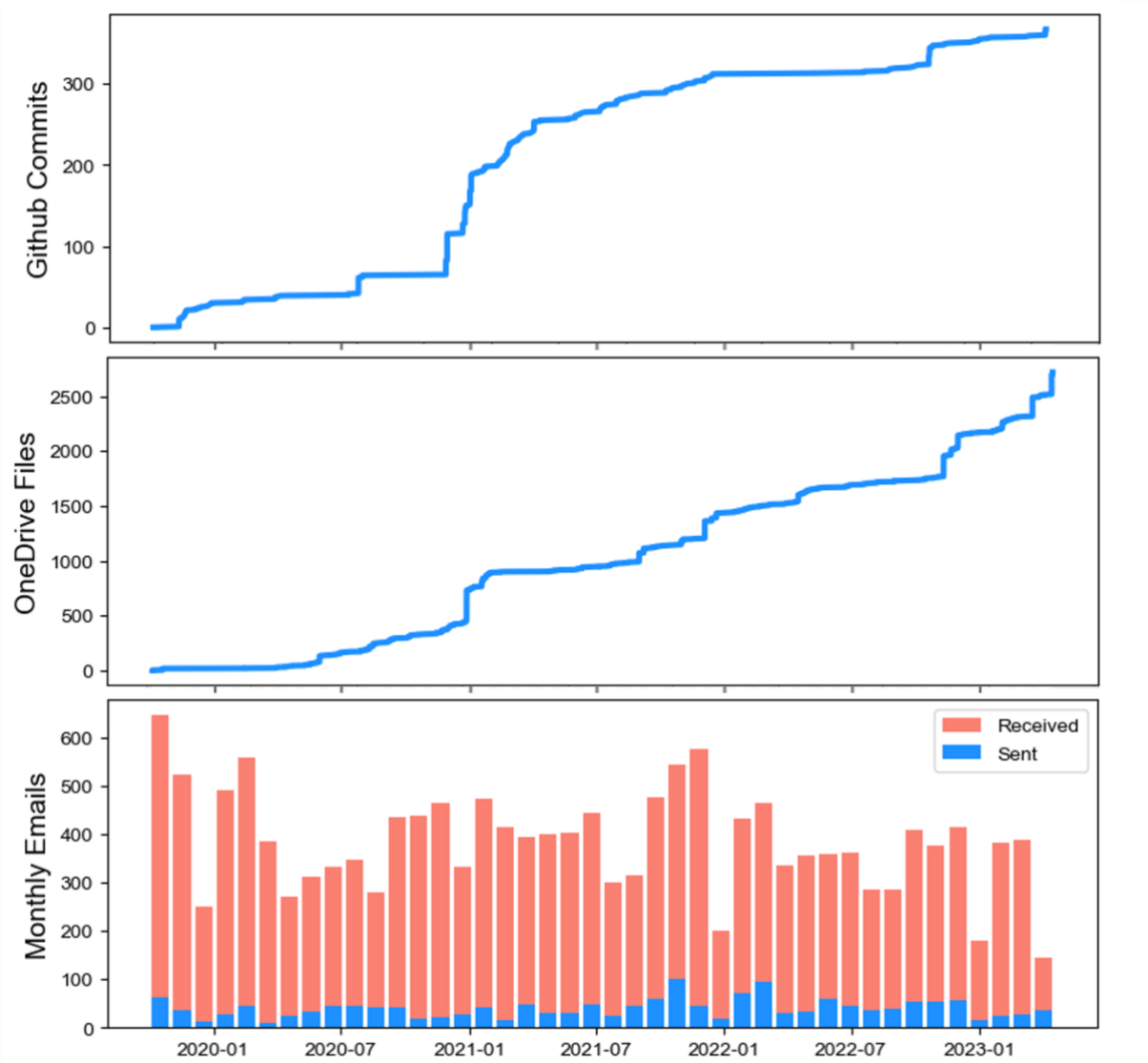

Figure 3 | Alternative productivity metrics. (Top) Cumulative number of code commits to PhD-related Github repositories. (Middle) Cumulative number of PhD-related files backed-up on OneDrive. (Bottom) Total number of PhD-related emails sent (blue) and received (pink) each month.

Suddenly in **Figure 3** we now see a very different picture to what the conventional metrics showed in **Figure 2**. Rather than an exponential upward increase in productivity in the last months of the PhD, these alternative metrics appear to show a relatively constant and persistent level of productivity. In fact, linear fits of the cumulative trends of these alternative metrics produce R^2 values of 0.84 0.98, and 0.98/0.99 for Github commits, OneDrive files, and sent/received emails respectively.

This suggests that the beginning of the PhD is not the unproductive, low-output time that one can become convinced of when considering conventional metrics, and nor is the end of the PhD a frenzy of enlightened, super-productive activity. Rather, this tells us that our levels of productivity are rather constant throughout the entire PhD. This brings to mind Vonnegut's discussion on the shape of stories from his unfinished master's thesis[2]; he argues the arc of many fiction stories follow one of just a few shapes (e.g. 'Man-in-hole': someone gets into trouble, gets out of it again…), but that the most realistic stories in life have a rather 'flat' trajectory, reflecting an unshakable ambiguity that accompanies life events. This seems to mesh well with the narrative of scientific research where we are often unsure as to whether our measurements are 'good' or 'bad', but in reality, they simply 'are'. As scientists perhaps the best we can do is to remain persistent.

Despite this 'flatness', there are still a couple of interesting points that can be picked out from **Figure *3***:

- A spike in Github commits at the end of 2020 reflects a large amount of code that resulted from my internship with Illumina; a slightly higher rate is seen to persist also for the following few months, likely because the same workstyle of high-frequency code uploads stuck with me when I returned to the PhD, or perhaps because I had nothing better to do during the third covid lockdown.
- On average I sent 40 ± 20 emails each month, whilst receiving 350 ± 90. Whilst that may sound like a lot, market research by Radicati suggests the average email account sends and receives a combined 80 emails *per day*[3]. I must therefore conclude that either PhD students are actually relatively well-insulated from administrative email pressures, or that the University of Cambridge has *really* good anti-spam filters.

The final metric that I explored was the word count of my weekly diary. Each week throughout my PhD, I journaled into a template with three sections: intentions for the week to come; notes and tasks for each day; and reflections looking back over the week. These varied significantly in length, from elaborate simulation plans with checkpointed to-do lists and comprehensive reviews of what to improve next week, to short plans such as "Write write write!", and succinct summaries like "Doom – tried to fix code but failed…". In **Figure *4*** the weekly word counts of these diary entries are plotted over the full course of the PhD as well as showing a four-week rolling average.

One of the major initial impressions in **Figure *4*** is the clear downward trend over time, with much higher wordcounts at the start of my PhD than by the end. Is this yet another contradiction? I didn't get more productive as suggested by citations, or remain constant as suggested by my frequency of file creation, but actually got less productive? Whilst this could be a possibility, I think perhaps it rather reflects a reduced need to rely on complicated plans. By the end of my PhD, I knew exactly what needed to be done and so plans and notes could be relatively minimal, whereas at the start those plans were crucial for helping me to scope out next steps and settle into the role of being a PhD student.

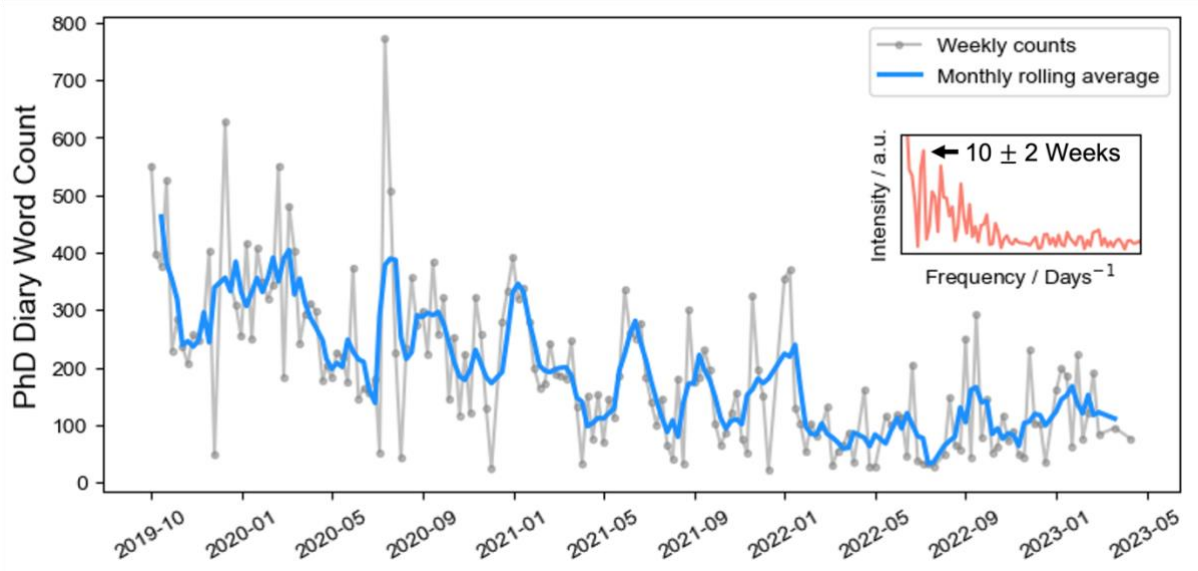

Figure 4 | Word count of weekly PhD diary. Word counts of weekly PhD diary (nominally containing weekly aims, daily notes/tasks, and weekly reflections). Grey data points show individual counts for each week whilst the blue line indicates a four-week rolling average. Inset shows absolute intensity of the Fourier-transformed signal, with major period identified (Frequency ranges from 0 to 0.035 days^{-1}).

It is also possible to pick out specific low and high points from ***Figure 4***. For example, the week of Christmas each year shows the wordcount drop near to zero (mostly with just 'template' words remaining), and other holidays are clearly visible as similar dips. At the higher end of the scale, a maximum peak can be seen in week 45 of the PhD (August 2020) – this, by no coincidence, is the week I received reviewer's comments on my first paper submission and was sent into overdrive attempting to make all the necessary changes for resubmission before going on holiday and leaving to work for Illumina in September. I recall at the time thinking the feedback was very negative (reviewers certainly don't give positive/negative feedback in the 'ideal' 5:1 ratio[4]), but in reality it was a very useful, constructive response that certainly led to an improved piece of work – something I only realised some months later.

One of the things I found most striking looking at the rolling average in ***Figure 4*** was the peaks and troughs that appear to oscillate at a roughly constant frequency. By performing a Fourier decomposition and analysing the strongest non-zero frequency, I determined that my diary wordcounts did indeed oscillate with a period of 10 $\pm$ 2 weeks, implying perhaps that I worked in productive bouts of 4 to 6 weeks, followed by a similar length dip in productivity. I find it absolutely fascinating to see this fall out of the data, as this work pattern was certainly not something I consciously aimed for. At first, I thought this could be to down to a 'refreshing' factor from that results from taking frequent holidays, but research shows holiday frequency actually has no significant impact on happiness levels[5], and so perhaps not on productivity levels either. But then I remembered, throughout all of my school years I also worked in 'periods' of approximately 10 weeks (each 'half-term' in the UK is typically 8 weeks followed by 1-2 weeks of holiday), could this be an unconscious influence? Furthermore, research into the ideal length for undergraduate courses found that intensive 4- and 8-week courses resulted in greater performance than the same content taught over a traditional 16-week semester[6]. Taking this all into consideration,

I think that perhaps the freedom and autonomy that came with the PhD enabled me to naturally find and fall into my most productive working pattern – even if I wasn't aware of it!

Conclusions

PhDs are a challenging but rewarding journey, and whilst they can be analysed through many metrics of performance, none of them quite tell the whole story. Students at the beginning of their PhD should not dwell on the quantity of their outputs, but can rather be confident that even if they don't feel like it, they are laying down important plans for the latter stages of the PhD. Similarly, activities such as teaching or working part-time jobs, or indeed resting after periods of intense work, need not be viewed as something which detracts from your research, but can in fact provide a crucial structure that helps PhD students to organise their time. Ultimately though, it is down to the individual to determine the shape of their own PhD.

Acknowledgements

I would like to thank my supervisors Prof. Midgley and Prof. Ringe for their guidance throughout my PhD, without them none of this would have been possible. I also acknowledge my funding from EPSRC NanoDTC Cambridge (No. EP/L015978/1). The code I wrote for extracting this data and generating figures can be found at https://github.com/grlewis333/PhD-Public .

References

1. Mannella, R. & Rossi, P. On the time dependence of the h-index. *J Informetr* **7**, 176–182 (2013).

2. Vonnegut, K. Fluctuations Between Good and Ill Fortune in Simple Tales. (University of Iowa, 1965). (Video summary at https://youtu.be/4_RUgnC1lm8)

3. The Radicati Group. *Email statistics report 2021-2025*. https://www.radicati.com/wp/wp-content/uploads/2020/12/Email-Statistics-Report-2021-2025-Executive-Summary.pdf (2021).

4. Losada, M. & Heaphy, E. The Role of Positivity and Connectivity in the Performance of Business Teams. **47**, 740–765 (2004).

5. Nawijn, J. Happiness Through Vacationing: Just a Temporary Boost or Long-Term Benefits? *J Happiness Stud* **12**, 651–665 (2011).

6. Austin, A. & Gustafson, L. Impact of course length on student learning. *Journal of economics and finance education* **5**, 26 (2006).