Grassroots Training for Reproducible Science: A Consortium-Based Approach to Enhancing Transparency and Rigor in Empirical Dissertation Research

Dr. Anna Keller[1]*, Dr. Lukas Meier[1]

[1]ETH Zurich, Department of Biomedical Data Science and Research Methodology, Zurich, Switzerland

Abstract

There is a widely acknowledged need to improve the reliability and efficiency of scientific research to increase the credibility of the published scientific literature and accelerate discovery. Widespread improvement requires a cultural shift in both thinking and practice, and better education will be instrumental to achieve this. Here we argue that education in reproducible science should start at the grassroots. We present our model of consortium-based student projects to train undergraduates in reproducible team science. We discuss how with careful design we have aligned collaboration with the current conventions for individual student assessment. We reflect on our experiences of several years running the GW4 Undergraduate Psychology Consortium offering insights we hope will be of practical use to others wishing to adopt a similar approach. We consider the pedagogical benefits of our approach in equipping students with 21st Century skills. Finally, we reflect on the need to shift incentives to reward to team science in global research and how this applies to the reward structures of student assessment.

Keywords: reproducibility, reproducible research, team science, collaboration, student projects, empirical dissertation, pre-registration, open science.

Grassroots training for reproducible science: a consortium-based approach to the empirical dissertation

Across the social and life sciences there has been a growing appreciation that many published findings may be unreliable (Begley & Ioannidis, 2015; Chalmers et al., 2014; Chan et al., 2014; Glasziou et al., 2014; Ioannidis et al., 2014; Salman et al., 2014). In one analysis, 85% of efforts in biomedical research were deemed to be wasted through inefficiency and poor methods (Macleod et al., 2014). In a recent survey in Nature, 90% of respondents agreed there was a "reproducibility crisis"(Baker, 2016). Small sample sizes and questionable research practices such as p-hacking, hypothesising after the results are known, and selectively publishing positive results were all implicated as contributing factors. Psychology has not been immune. Classic findings have failed to replicate and experimental findings have failed to translate into effective interventions (Cristea, Kok, & Cuijpers, 2015; Doyen, Klein, Pichon, & Cleeremans, 2012; Ritchie, Wiseman, & French, 2012). In 2015, the Open Science Collaboration reported a replication rate of 39% for the 100 effects they attempted to replicate (Aarts et al., 2015).

In response there is a move towards rigorous and transparent scientific practices, such as pre-registering study protocols and making materials and data open-access; designing studies with sufficient statistical power; and team-science approaches to pool resources and expertise (Munafò et al., 2017). Until these ‘new’ methods become the norm, however, the extra time and resources required for bigger samples and pre-registration, the reduced probability for leveraging chance findings, and the lack of conventions to adequately reward collaborative efforts may act as disincentives. Indeed, many, including ourselves, have argued that our current reward structures are failing to reward good science (Button et al., 2013; Button & Munafo, 2013; Munafò et al., 2017; Nosek, Spies, & Motyl, 2012).

Improved education coupled with shifting incentives are repeatedly identified as the key levers to promote better practices to address the 'reproducibility crisis' (Munafò et al., 2017).

When asked what factors could boost reproducibility, better training and teaching followed by incentives for better practice were ranked top by the respondents in the Nature survey (Baker, 2016). While there are moves to change the supra-order incentives structures researchers work within, these may take time to embed. For lower-order initiatives such as improved methods and education to be effective they will need to work alongside, rather than orthogonal to, the current reward and recognition structures.

We suggest that education in reproducible science should start at the grassroots. We present our model of consortium-based student projects to train undergraduates in reproducible team science. We discuss how our approach is carefully designed to align reproducible research methods training with the current conventions for individual student assessment, aligning quality teaching with the production of quality research.

The wider "reproducibility crisis": causes and solutions

The factors contributing to the reproducibility crisis are multi-layered, interacting and complex. The solutions will need to be correspondingly multifaceted, each focusing on a specific element whilst also considering the wider eco-system. In the context of improving education, it is important to target the main areas which perpetuate the current issues. These include those operating at the individual study level, such as poor experimental design, the individual researchers' behaviours, such as selective decisions of what and what not to publish, and the wider recognition structures in academic culture. We will discuss these briefly alongside some of the recommended solutions before discussing how we implement these solutions in our undergraduate training.

Low statistical power

Poor study design undermines the reliability of the subsequent study results. Studies consistently find statistical power to be low, between 8~50%, across both time and disciplines studied (Button et al., 2013; Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Low statistical power increases the likelihood of obtaining both false positive and false negative

results as well as generating inflated effect size estimates (Button et al., 2013). A reliance on under-powered studies therefore undermines the accumulation of knowledge.

Solutions include better education in the principles of sound research design. Performing sample size calculations is crucial to determining the appropriate study design, however very few studies actually report a justification of their sample size (Bahor et al., 2017), and scientists seem overly optimistic about the likely size of effects under test. Researchers could use effect sizes from the existing literature, adjusted for publication bias, as the basis of their sample size calculations, or more work could be done around minimal clinically or theoretically important differences to inform sample size calculations (Angst, Aeschlimann, & Angst, 2017; Button et al., 2015; Jaeschke, Singer, & Guyatt, 1989).

Low powered research persists due to perverse incentives and a poor understanding of the consequences of low power and /or a poor appreciation of likely effect sizes. However, it may also arise from a lack of resources to improve power. Team science is a solution to the latter problem. Instead of relying on the limited resources of single investigators, distributed collaboration across many study sites could enable high-powered designs. Multi-site testing also confers advantages in the generalisability of results across settings and populations sampled, the transfer of knowledge through bringing multiple theoretical and disciplinary perspectives, and the benefits of a diverse range of research cultures and experiences on the same project.

Multi-centre large-scale collaborations have a long and successful tradition in fields such as randomized controlled trials, and in genetic association analyses, and have improved the robustness of the resulting research literatures. Multi-site collaborative projects have also been advocated for other types of research, such as animal studies (Bath, Macleod, & Green, 2009; Dirnagl et al., 2013; Milidonis, Marshall, Macleod, & Sena, 2015), in an effort to maximize their power, enhance standardization, and optimize transparency and protection from biases. The Many Labs and Pipeline projects illustrate this potential in the social and

behavioural sciences, with dozens of laboratories implementing the same research protocol to obtain highly precise estimates of effect sizes, and evaluate variability across samples and settings (Ebersole et al., 2016; Klein et al., 2014; Schweinsberg et al., 2016).

Publication bias and questionable research practices

Studies that obtain positive and novel results are more likely to be published than studies that obtain negative results or report replications of prior results (Franco, Malhotra, & Simonovits, 2014; Sterling, 1959, 1975). This is known as publication bias (Sterling, 1959). Over 70% of authors admit to selectively reporting studies that "worked" – with those that failed to find significant results languishing in a file drawer (John, Loewenstein, & Prelec, 2012).

Questionable research practices refer to a body of behaviours researchers employ to leverage chance to increase the likelihood of a publishable result. Examples include, measuring multiple outcomes and switching the 'primary outcome' to whichever produces the strongest effect. P-hacking, that is running multiple variations of analyses until a significant p-value is obtained (Simmons, Nelson, & Simonsohn, 2011), and Hypothesising After the Results are Known (HARKing), where a hypothesis is retro-fitted to an unexpected result (Kerr, 1998). Such behaviours increase the probability of type I errors and undermine the reliability of results.

Pre-registration is the most powerful way to protect against all of these practices. It was introduced in clinical trials to address publication bias and outcome switching where it has now been standard practice for many years (Dickersin & Rennie, 2003; Lenzer, Hoffman, Furberg, Ioannidis, & Grp, 2013). In its simplest form it registers the basic study design and primary outcome. In its strongest form, it involves registering the study with a commitment to make the results public, and detailed pre-specification of the study design, primary outcome, and analysis plan in advance of data-collection. This makes the research discoverable, and addresses questionable research practices such as outcome switching and P-hacking. More

fundamentally, it allows a clear distinction between data-independent confirmatory research that is important for *testing* hypotheses, and data-contingent exploratory research that is important for *generating* hypotheses. Websites such as the Open Science Framework (http://osf.io/) and AsPredicted (http://AsPredicted.org/) offer services to pre-register studies, and over 150 journals now adopt the Registered Reports publishing format (Chambers, 2013; Nosek & Lakens, 2014).

Perverse incentives

Poor experimental design, small sample sizes, undisclosed flexibility in data analysis and a lack of transparency in reporting all undermine the purpose of the scientific endeavour (assuming the purpose is to generate knowledge). Why then do they persist? Publication is the currency of a successful academic career increasing the likelihood of employment, funding, promotion, and tenure. However, as we have discussed above, positive, novel, clean results are more likely to be published than negative results, replications, or 'messy' results (Dickersin, 1990). Consequently, researchers are incentivized to focus on producing positive results irrespective of their reliability (Nosek, Spies, & Motyl, 2012; Smaldino & McElreath, 2016). Researcher behaviour in the context of these perverse incentives has been modelled mathematically using simulations. Worryingly, these models suggest that researchers acting to maximise their “fitness” in the current incentive structures should spend most of their effort seeking novel results and conduct small studies that have only 10-40% statistical power (Higginson & Munafo, 2016).

Student projects: the problem magnified

This perverse reward culture may be operating at the first stages of training, in how we teach and assess our undergraduate students. For example, in undergraduate psychology degrees the UK, the culmination of a student's research training is their final year empirical dissertation (a research project akin to the senior thesis in the US system). It involves carrying out an extensive piece of empirical research, under the supervision of an assigned academic (much like the PhD supervision model). It requires the student to individually

demonstrate a range of research skills including planning, considering and resolving ethical issues, and the analysis and dissemination of findings (BPS, 2016). This empirical project typically involves the collection of original data, or equivalent alternatives such as secondary data analysis, and forms a substantial proportion of the student's final grade (BPS, 2016).

This leads to numerous final year research projects, individually conducted and assessed, and conducted with limited time and resources. Consequently, they may suffer from the problems seen in the wider research literature, such as small sample sizes, combined with questionable research practices which may yield a high rate of false positive results. If these are selectively published, the students will be rewarded for being lucky rather than right. Indeed, an undergraduate publication is seen as highly competitive when applying for places on taught-Masters or for doctoral training.

As an example, suppose each year a psychology department has 100 final year undergraduate students conducting a quantitative dissertation. Each student has been encouraged to ‘think creatively’ and generate a novel, exciting hypothesis. Their hypothesis might therefore be quite unlikely to be true. Indeed, let us assume that only 10 / 100 students’ hypotheses are ‘true’. Now let us assume they are limited in time and resources and their samples sizes are correspondingly small for the effects under test, yielding an average statistical power of 20%. They set their significance criterion for rejecting the null hypothesis at 5%. Under these assumptions 20% of the 10 ‘true’ effects will be discovered, yielding 2 true positive results. For the remaining 90 non-associations, 5% will be declared as significant by chance, yielding on average between 4 and 5 false positive results. Now let us assume supervisors encourage students to write up any interesting positive findings for publication, two-thirds of those will be false-positive. This is known as the positive predictive value (Sterne & Davey Smith, 2001). This would contaminate the scientific literature with unreliable results.

Developing a comprehensive pre-registration often takes several months, and unless effects under test are very large, a sample size of 200-300 hundred participants may be required for sufficient power to test typical effects observed in psychology. Pre-registration of study protocols and designing studies with sufficient power to test hypotheses will therefore often require resources well beyond those available for the typical undergraduate dissertation project. Drawing on the above examples of large-scale collaborations in genetic epidemiology and multi-centre clinical trials, one solution to this is collaboration. By working together, students could pool their efforts and resources to obtain large data-sets. Any limitations imposed by the need for individual students' assessment within an institution could be overcome by students working across universities.

The GW4 Undergraduate Psychology Consortium

In 2016 we set up the GW4 Undergraduate Psychology Consortium, comprising academics and undergraduates, and more recently PhD students and post-doctoral researchers, from the University of Bath, University of Bristol, Cardiff University and Exeter University, in the south-west of the UK (Button, Lawrence, Chambers, & Munafo, 2016). The composition of the consortium varies year-to-year but the basic structure is provided in Figure 1. Students work in groups with-in and across the universities under the supervision of their local academic supervisor, and PhD student or post-doctoral researcher. Each year the consortium focuses on a single research project with a primary research question lead by one of the intuitions. To ensure fairness, the lead institution rotates year-to-year sharing the rewards and responsibilities. The research cycle is typically two years. Rotating the lead ensures no single institution is over-burdened.

Assigning an individual study lead, or principal investigator, is essential for the effective coordination of the study. Given the extra work involved in pre-registration, coordination of multiple sites, and guaranteeing provision for the analysis and write-up of the project for publication after the undergraduate students have graduated, we have found it works best if the lead is a PhD student or post-doctoral researcher. The consortium benefits from

additional support and dedicated oversight and this frees up the academics to focus on academic supervision. The lead PhD student or post-doctoral researcher benefits from gaining experience of leading a collaboration under the supervision of more experienced academics, and of mentoring undergraduate students. They also benefit from a first-author publication.

The empirical dissertation timeline and allocation procedures vary from University to University, and this is one of the biggest challenges in the consortium approach. Typically, students are assigned (with varying degrees of student choice) to their project supervisor anywhere from February through to September in their penultimate year of study. Information outlining the objectives and general procedure, including the home-work element in the weeks prior to the start of term, are shared with the students during allocation. The final year dissertation project then commences in October through to the dissertation deadline which ranges from early April through to late May.

From the preceding January through to June the lead PhD student or post-doctoral researcher and academic supervisory team develop a primary research question and bare-bones study protocol (containing hypotheses, rationale and background, methods, data analysis and management plan, ethical considerations). A few weeks before the start of term this is shared with the upcoming cohort of undergraduate students who are tasked in a home-work exercise to think about what additional secondary questions they might like to add to the protocol. These secondary questions range from moderator analyses based on adding short individual difference measures, to focusing on a different outcome, or using the variables already in the design to address a different question. Each student, supported by their local supervisor, is instructed to prepare a 5-minute presentation outlining their proposed secondary research question, background and rationale, hypothesis, measures and methods, and analysis plan, including a graph of their hypothesised result.

In the first or second week of term in October we have the first of two consortium meetings. During this meeting the rationale for the consortium approach is presented, followed by the lead PhD student or post-doctoral researcher presenting an overview of the primary research question and procedure. Each student then presents their proposed secondary hypothesis. This is followed by general discussion of merits of each of the proposed hypotheses and corresponding measures and methods. The study protocol is then finalised through these collaborative discussions and all students are authors of the published study protocol. For example, suppose two students suggest a similar hypothesis, that the effect of X on Y is moderated by Z, but have proposed different methods of measuring Z. There is a clear issue of redundancy, and which measure to include is decided by considering issues such as psychometric properties, relevance to the population of interest, and so forth.

Following the first meeting the students are assigned various tasks relating to study set-up; such as constructing the surveys for data-collection, piloting the experimental tasks and procedures, selecting stimuli, and drafting ethics applications, case-report forms, information sheets, debrief sheets and recruitment advertisements. The project is managed centrally on the Open Science Framework (http://osf.io/).

The study protocol is pre-registered on the Open Science Framework and data collection runs from late November through to the end of February / March. Each site has a target sample size corresponding to a proportion of the total sample size calculated for the overall project, adjusted for their institutional requirements and recruitment challenges. Data collection finishes at the end of February / March to accommodate the students with the earliest dissertation deadlines.

In March / April we hold the second consortium meeting, akin to a mini-conference, where each student presents their dissertation results. There is a general discussion about the study findings and the general conclusions of the final publication are mutually agreed, thus

each student qualifies for authorship on any publications. To date, students have used the full data-set from across the consortium in their dissertations.

Aligning rigorous research with individual assessment and programme standards

In the UK Psychology degrees are accredited by the British Psychological Society (BPS; the equivalent of the APA in the US) which sets the programme standard. Accredited psychology degrees include the acquisition of a range of research skills and methods, culminating in an ability to conduct research independently. Graduates must demonstrate competency in designing, conducting and reporting on empirically-based research projects under appropriate supervision (BPS, 2016). There are various opportunities for developing these skills through the degree, culminating in the final year in the empirical dissertation (BPS, 2016). The GW4 Psychology Consortium was designed specifically with this degree standard in mind. By collaborating, students benefit from pooled resources and data collection. The pre-registered primary research question and hypothesis means the integrity of the overall research project is retained. The addition of student-led secondary hypotheses means each student can have a different title and focus of their dissertation, so with-in an institution no two students are doing exactly the same thing. Thus throughout, each student can demonstrate individual contributions to each element of the research but within the scaffold of a high-quality research project. In 2016, we presented the GW4 consortium approach to the BPS Undergraduate Education Committee, and they revised their accreditation guidance around the empirical dissertation to explicitly endorse group-based projects, provided students are able to demonstrate the required research skills individually (BPS, 2016).

Pedagogy

There is broad consensus amongst educators, business leaders, academics, and governmental agencies, that traditional academic skills that have focused on knowledge-based content and surface-learning are no longer sufficient to prepare our students for success in the 21st Century. Instead, they have identified skills associated with deeper

learning, such as analytic reasoning, complex problem solving, and teamwork as crucial for meeting the demands of the modern-day workforce (OECD, 2005). In science, collaboration is increasingly recognised by funders and key stakeholders as necessary to meet global challenges (Academy of Medical Sciences, 2016). The number of authors on the typical research paper increases year-on-year, and the typical lab-group runs as a team (Academy of Medical Sciences, 2016). For psychology to be well placed to meet these challenges we must find ways to train and support, and then adequately recognize and reward, psychologists who contribute to team efforts, starting with undergraduate education.

Our consortium approach to the undergraduate empirical dissertation has many pedagogical advantages. It is project-based learning grounded in social-constructivism (Roessingh & Chambers, 2011). Students join a knowledge-generating community, where, in collaboration with others, they conduct real research in an authentic context. They gain the transferrable skills of collaboration, problem-solving and creating new knowledge through co-production. Through peer-to-peer interaction, they actively apply their knowledge in discussion, particularly during the consortium meetings, but also in the day-to-day conduct of the study.

The consortium structure creates “zones of proximal development” where students contribute to each element of the research cycle, but within a scaffold which retains the integrity of the overall research. The 'bare-bones' protocol drafted by the lead PhD student, is an example of this. It provides the students with a scaffold of best-practice which they can model in their preparation of their secondary hypothesis and corresponding background, method, and analysis plan.

One limitation of this approach is the scope for "social loafing" (Karau & Williams, 1993; Latane, Williams, & Harkins, 1979). This might occur where some members of the team are not as intrinsically motivated to work as hard on collaborative aspects, such as data collection, as others. For example, knowing that they will have access to the whole data set regardless of the number of participants that they themselves recruit, might demotivate

some individuals. Social loafing may be more likely where assessment and/or reward structures do not recognise team work. For example, where a group is awarded a common mark regardless of individual contributions (Lee & Lim, 2012), or in terms of the dissertation, where the final grade reflects the write-up, and ignores the student's teamwork contributions.

This reflects a wider academic debate surrounding the need for better recognition of individual contributions to team efforts. Following recommendations in the recent "Improving recognition of team science contributions in biomedical research careers" (Academy of Medical Sciences, 2016) the consortium members could collectively agree milestones, credit and responsibilities during the first consortium meeting. Contributions could be logged as the project progresses, and the student and/or supervisor could submit a statement of contribution alongside their individual dissertation (this is already standard practice at some Universities). Supervisors can also promote a culture of collaboration by actively acknowledging good teamwork in their day-to-day mentoring and sign-posting where collaboration will be rewarded beyond the dissertation grade. For example, in the UK, dissertation supervisors are often asked to provide an academic reference. This provides an opportunity to acknowledge the student's contribution to the wider team effort, providing evidence of both academic achievement in the dissertation grade, and softer 21st Century skills in appraisal of the student's conduct during the study.

Evaluation

To date, 24 students have participated in and successfully completed their dissertation project as part of the GW4 Consortium. The consortium has published three pre-registrations (Button et al 2016, Hobbs et al 2019, Tzavella, Adams, Chambers, Lawrence, Button 2018), and has three manuscripts in preparation. Anecdotal evidence suggests that the scaffolding provided by the consortium approach helped the weaker students, and enabled the stronger students to excel. We plan to test this quantitatively by comparing grades awarded for the dissertation for consortium versus cohort students, adjusted for

grades awarded from non-dissertation units, when a sufficient number of students have completed to ensure their anonymity. We have been surveying the students at the end of each project for their feedback. They greatly valued participating in the consortium, particularly having access to a meaningfully large data-set, the opportunity to network with academics, students and researchers from other universities, the sharing of ideas and knowledge, and contributing to pre-registration. They saw the disadvantages as trying to coordinate deadlines and workload with other universities and having less control over the study design.

Wider applications and other collaborative initiatives in student projects

The consortium model we have presented here is just one example of an approach tailored to our way of working, our departmental courses and our national accreditation requirements for the empirical dissertation. Our consortium model would translate well to other educational systems which involve quantitative students' projects which collect data to test hypotheses, where students must individually demonstrate competency in design, conduct and reporting, but the timescale and resources mean a student would struggle on their own to collect sufficient data. For example, in the US the APA recommends capstone experiences for undergraduate training, which can include projects similar to the UK quantitative dissertation, and many undergraduate programs have either required or elective projects that are similar in size and scope. However, a core element of our approach is the two consortium meetings, which provide the opportunity for the students to independently contribute ideas to the study design, and any final publication. The GW4 Universities are in relatively close proximity to each other which means we can hold these meetings face-to-face in an afternoon. More geographically distributed collaborations may need to use conference calling, or other means to share ideas, which may lose some of the advantages for the students of networking in person with colleagues from other institutions. We have found that the protocol can become unwieldy if too many secondary hypotheses are added, so this limits the number of students from a single institution (our range is 2-6). However, this will depend on

the specific project design and institutional policy of how many students can address the same research question within an institution.

In undergraduate psychology methods teaching more broadly there are lots of exciting new initiatives. For example, The Collaborative Replications and Education Project (CREP; https://osf.io/wfc6u/) is a large-scale replication project that students can participate in. A coordinating team identifies recently published research that could be replicated in the context of a semester-long undergraduate research methods course. A central commons provides the materials and guidance to incorporate the replications into projects or classes, and the data collected across sites are aggregated into manuscripts for publication. This approach is particularly suited large-group methods teaching, which in the UK model typically occurs in the first and second years of the degree. It is less well suited to the final year empirical dissertation in the UK model, where the students must demonstrate individual contributions to each element of the research, and work closely with an academic supervisor in a manner more akin to the relationship between a PhD student and their supervisor. The “Pipeline” (Schweinsberg, Madan, Vianello, Sommer, Jordan, Tierney, Awtrey, Zhu, Diermeier, Heinze, Srinivasan, Tannenbaum, Bivolaru, Dana, Davs-Stober, et al., 2016) and “Many Labs” (Ebersole et al., 2016; Klein et al., 2014) and Psychological Science Accelerator (https://psysciacc.org/) projects also offer opportunities to contribute to large-scale replication efforts with coordinated data collection across many locations simultaneously.

Conclusions

The need to improve the reliability and efficiency of scientific research is widely acknowledged. Improved education is repeatedly identified as a key lever to meet this need. Consortium-based student projects are one example of how reproducible research methods training can be aligned with the current conventions for individual student assessment, and the production of high-quality research. The consortium approach has many pedagogical

benefits, equipping students with the 21st Century skills of collaboration, analytic reasoning, and complex problem solving. However, for these skills to be fully realised the assessment and marking criteria for the dissertation need to extend beyond individual assessment criteria to recognise and reward collaboration.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., . . . Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:ARTN aac471610.1126/science.aac4716

Angst, F., Aeschlimann, A., & Angst, J. (2017). The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. *J Clin Epidemiol, 82*, 128-136. doi:10.1016/j.jclinepi.2016.11.016

Bahor, Z., Liao, J., Macleod, M. R., Bannach-Brown, A., McCann, S. K., Wever, K. E., . . . Sena, E. (2017). Risk of bias reporting in the recent animal focal cerebral ischaemia literature. *Clinical Science, 131*(20), 2525-2532. doi:10.1042/Cs20160722

Baker, M. (2016). IS THERE A REPRODUCIBILITY CRISIS? *Nature, 533*(7604), 452-454. doi:10.1038/533452a

Bath, P. M. W., Macleod, M. R., & Green, A. R. (2009). Emulating multicentre clinical stroke trials: a new paradigm for studying novel interventions in experimental models of stroke. *International Journal of Stroke, 4*(6), 471-479. doi:10.1111/j.1747-4949.2009.00386.x

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in Science Improving the Standard for Basic and Preclinical Research. *Circulation Research, 116*(1), 116-126. doi:10.1161/circresaha.114.303819

BPS (2016). Standards for the accreditation of undergraduate, conversion and integrated Masters programmes in psychology. Retrieved from https://www.bps.org.uk/sites/bps.org.uk/files/Accreditation/Undergraduate%20Accreditation%20Handbook%202019.pdf

Button, K. S., Chambers, C., Lawrence, N., Adams, R. C., Morrison, S., Smith, A., … Coombs, E. (2016, November 23). GW4 Undergraduate Psychology Consortium 2016/17. Retrieved from osf.io/wxyra

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci, 14*(5), 365-376. doi:10.1038/nrn3475

Button, K. S., Kounali, D., Thomas, L., Wiles, N. J., Peters, T. J., Welton, N. J., . . . Lewis, G. (2015). Minimal clinically important difference on the Beck Depression Inventory--II according to the patient's perspective. *Psychol Med, 45*(15), 3269-3279. doi:10.1017/S0033291715001270

Button, K. S., Lawrence, N. S., Chambers, C. D., & Munafo, M. R. (2016). Instilling scientific rigour at the grassroots. *Psychologist, 29*(3), 158-159.

Button, K. S., & Munafo, M. R. (2013). Incentivising reproducible research. *Cortex*. doi:10.1016/j.cortex.2013.09.011

Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gulmezoglu, A. M., . . . Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *Lancet, 383*(9912), 156-165. doi:10.1016/s0140-6736(13)62229-1

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609-610. doi:10.1016/j.cortex.2012.12.016

Chan, A. W., Song, F. J., Vickers, A., Jefferson, T., Dickersin, K., Gotzsche, P. C., . . . van der Worp, H. B. (2014). Increasing value and reducing waste: addressing inaccessible research. *Lancet, 383*(9913), 257-266. doi:10.1016/s0140-6736(13)62296-5

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *J Abnorm Soc Psychol, 65*, 145-153.

Cristea, I. A., Kok, R. N., & Cuijpers, P. (2015). Efficacy of cognitive bias modification interventions in anxiety and depression: meta-analysis. *British Journal of Psychiatry, 206*(1), 7-16. doi:10.1192/bjp.bp.114.146761

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA, 263*(10), 1385-1389.

Dickersin, K., & Rennie, D. (2003). Registering clinical trials. *JAMA, 290*(4), 516-523. doi:10.1001/jama.290.4.516

Dirnagl, U., Hakim, A., Macleod, M., Fisher, M., Howells, D., Alan, S. M., . . . Meairs, S. (2013). A Concerted Appeal for International Cooperation in Preclinical Stroke Research. *Stroke, 44*(6), 1754-1760. doi:10.1161/strokeaha.113.000734

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral Priming: It's all in the Mind, but Whose Mind? *PLoS One, 7*(1). doi:ARTN e2908110.1371/journal.pone.0029081

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J Exp Soc Psychol, 67*, 68-82.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*, 1502-1505.

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., . . . Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *Lancet, 383*(9913), 267-276. doi:10.1016/s0140-6736(13)62228-x

Higginson, A. D., & Munafo, M. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biol, 14*(11), e2000995.

Hobbs, C., Al Balushi, R., Antony, P., Jordan, D., Lee-Hannah, M., … Button, K. S. (2018, April 1). Protocol. Compete or Cooperate? The impact of social context on self-other performance estimates https://doi.org/10.17605/OSF.IO/BNZDE

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., . . . Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *Lancet, 383*(9912), 166-175. doi:10.1016/s0140-6736(13)62227-8

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials, 10*(4), 407-415.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychol Sci, 23*(5), 524-532. doi:10.1177/0956797611430953

Joint training project : awareness pack. [Leeds]: NHS Executive.

Karau, S. J., & Williams, K. D. (1993). Social Loafing - a Metaanalytic Review and Theoretical Integration. *J Pers Soc Psychol, 65*(4), 681-706. doi:Doi 10.1037/0022-3514.65.4.681

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev, 2*(3), 196-217. doi:10.1207/s15327957pspr0203_4

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. J., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Soc Psychol, 45*, 142-152.

Latane, B., Williams, K., & Harkins, S. (1979). Many Hands Make Light the Work - Causes and Consequences of Social Loafing. *J Pers Soc Psychol, 37*(6), 822-832. doi:Doi 10.1037//0022-3514.37.6.822

Lee, H. J., & Lim, C. (2012). Peer Evaluation in Blended Team Project-Based Learning: What Do Students Find Important? *Educational Technology & Society, 15*(4), 214-224.

Lenzer, J., Hoffman, J. R., Furberg, C. D., Ioannidis, J. P. A., & Grp, G. P. R. W. (2013). Ensuring the integrity of clinical practice guidelines: a tool for protecting patients. *Bmj-British Medical Journal, 347*. doi:ARTN f553510.1136/bmj.f5535

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P., . . . Glasziou, P. (2014). Biomedical research: increasing value, reducing waste. *Lancet, 383*(9912), 101-104. doi:10.1016/S0140-6736(13)62329-6

Milidonis, X., Marshall, I., Macleod, M. R., & Sena, E. S. (2015). Magnetic Resonance Imaging in Experimental Stroke and Comparison With Histology Systematic Review and Meta-Analysis. *Stroke, 46*(3), 843-+. doi:10.1161/strokeaha.114.007560

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. doi:10.1038/s41562-016-0021

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A method to increase the credibility of published results. *Soc Psychol, 45*, 137-141.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012b). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science, 7*(6), 615-631. doi:10.1177/1745691612459058

OECD (2005). *The Definition and Selection of Key Competencies*. Retrieved from http://www.oecd.org/pisa/35070367.pdf

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS One, 7*(3). doi:ARTN e3342310.1371/journal.pone.0033423

Roessingh, H., & Chambers, W. (2011). Project-Based Learning and Pedagogy in Teacher Preparation: Staking Out the Theoretical Mid-Ground. *International Journal of Teaching and Learning in Higher Education, 23*(1), 60-71.

Salman, R. A., Beller, E., Kagan, J., Hemminki, E., Phillips, R. S., Savulescu, J., . . . Chalmers, I. (2014). Increasing value and reducing waste in biomedical research regulation and management. *Lancet, 383*(9912), 176-185. doi:10.1016/s0140-6736(13)62297-7

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., . . . Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55-67. doi:10.1016/j.jesp.2015.10.001

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., . . . Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *J Exp Psychol Gen, 66*, 55-67.

Sciences, A. o. M. (2016). *Improving recognition of team science contributions in biomedical research careers*. Retrieved from https://acmedsci.ac.uk/file-download/38721-56defebabba91.pdf

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychol Bull, 105*(2), 309-316.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci, 22*(11), 1359-1366. doi:10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *R Soc Open Sci, 3*(9), 160384. doi:10.1098/rsos.160384

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *J Am Stat Assoc, 54*, 30-34.

Sterling, T. D. (1975). Consequence of prejudice against the null hypothesis. *Psychol Bull, 82*, 1-20.

Sterne, J. A., & Davey Smith, G. (2001). Sifting the evidence-what's wrong with significance tests? *BMJ, 322*(7280), 226-231.

Tzavella, L., Adams, R. C., Chambers, C., Lawrence, N., & Button, K. S. (2018, June 1). The effect of inhibition training on approach-avoidance tendencies for unhealthy foods. Retrieved from osf.io/hz2nb

Figure 1. The typical structure of the GW4 Undergraduate Psychology Consortium

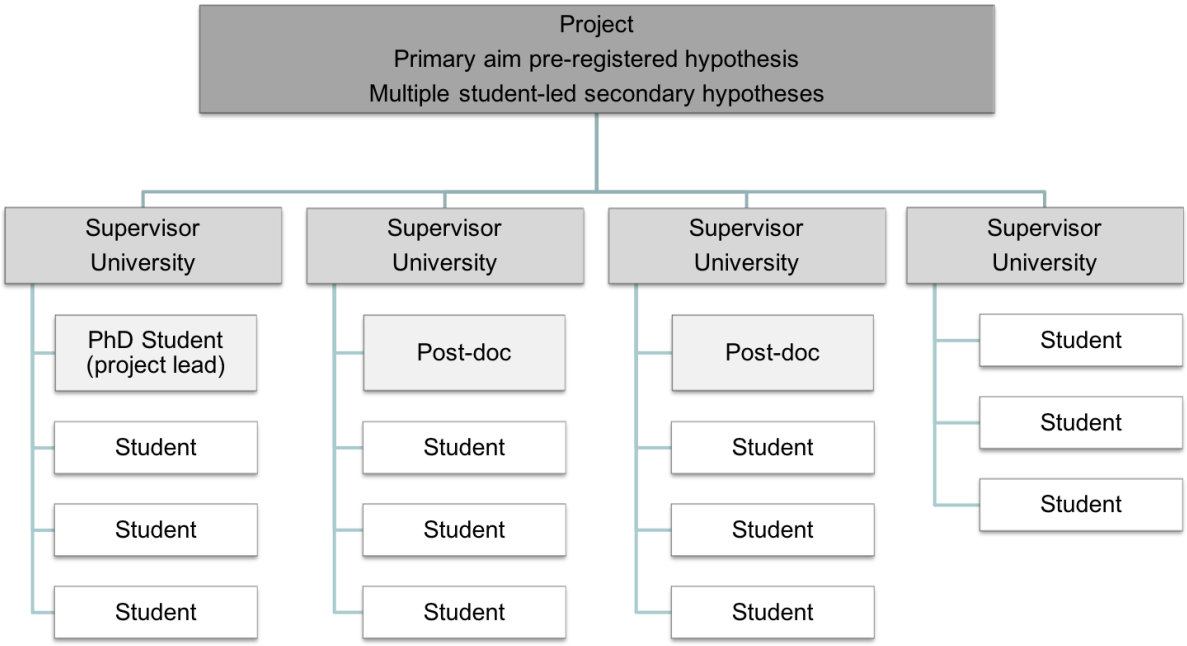

Figure 2. The typical timeline of the GW4 Undergraduate Psychology Consortium