Article:
Watson, M., Christoforou, P., Herrera, P. et al. (20 more authors) (2019) An analysis of the quality of experimental design and reliability of results in tribology research. Wear, 426-427 (Part B). pp. 1712-1718. ISSN 0043-1648

https://doi.org/10.1016/j.wear.2018.12.028

An analysis of the quality of experimental design and reliability of results in tribology research

Michael Watson*, Peter Christoforou, Paulo Herrera, Daniel Preece, Julia Carrell, Matthew Harmon, Peter Krier, Stephen Lewis, Raman Maiti, William Skipper, Ellis Taylor, Jonathan Walsh, Mohanad Zalzalah, Lisa Alhadeff, Reuben Kempka, Joseph Lanigan, Zing Siang Lee, Benjamin White, Kei Ishizaka, Roger Lewis, Tom Slatter, Rob Dwyer-Joyce, Matthew Marshall

Department of Mechanical Engineering, The University of Sheffield, Mappin Street, Sheffield, U.K. S1 3JD

Abstract

In recent years several high profile projects have questioned the repeatability and validity of scientific research in the fields of psychology and medicine. In general, these studies have shown or estimated that less than 50% of published research findings are true or replicable even when no breaches of ethics are made. This high percentage stems from widespread poor study design; either through the use of underpowered studies or designs that allow the introduction of bias into the results.

In this work, we have aimed to assess, for the first time, the prevalence of good study design in the field of tribology. A set of simple criteria for factors such as randomisation, blinding, use of control and repeated tests has been made. These criteria have been used in a mass review of the output of five highly regarded tribology journals for the year 2017. In total 379 papers were reviewed by 26 reviewers, 28% of the total output of the journals selected for 2017.

Our results show that the prevalence of these simple aspects of study design is poor. Out of 290 experimental studies, 2.2% used any form of blinding,

*Corresponding author
Email address: mea08mw@sheffield.ac.uk ()

3.2% used randomisation of either the tests or the test samples, while none randomised both. 30% repeated experiments 3 or more times and 86% of those who repeated tests used single batches of test materials. 9.2% completed any form of statistical test on their data.

Due to the low prevalence of repeated tests and statistical analysis it is impossible to give a realistic indication of the percentage of the published works that are likely to be false positives, however these results compare poorly to other more well studied fields. Finally, recommendations for improved study design for researchers and group design for research group leaders are given.

Keywords: Tribology, Replication, Experimental design,

2018 MSC: 00-01, 99-00

---

1. Introduction

In recent years several high profile projects and publications have questioned the repeatability of scientific research in general. One notable replication study in psychology [1] aimed to replicate the results of one hundred papers published in highly regarded journals of the field. This study failed to reproduce the original results of two thirds of the sample. Ioannidis [2] made estimates for the percentage of findings in the medical literature that were likely to be true, the study concluded that less than 50% are expected to be true for most experimental designs.

There are many reasons for erroneous results making it to publication. Most obviously it is possible that random chance can produce a result that seems important when no real effect is present. This is more likely when sample sizes are smaller, effect sizes are smaller [3], when there is a greater number and less pre-selection of tested relationships [4] and when statistical thresholds are too low [2, 5]. This problem is exacerbated by the bias to publish positive results and leave negative results unpublished [6, 7].

Bias can also be introduced into the design of the experiment and the data analysis methods used. This is not always obvious; for example if a near sig-

nificant result is found, adding further repeats to the experiment to push for a significant result increases the false positive rate [8]. This and other seemingly 'border line' methods, such as poor normalisation, have been used deliberately to show erroneous results including that listening to children's music makes the listener younger (p=0.03) [8].

These problems have driven some fields of science to stricter rules for publishing. In physics and genomic studies the threshold p values for statistical tests have been lowered, this case has also been made in other areas of science [5]. Many medical journals now require pre-registration of a trial, including data analysis techniques, before the trial starts, on the understanding that the trial will be published regardless of the outcome. Kaplan et al. [7] found that this reduced the proportion of positive results in one heart disease journal from 57% to 8%.

These problems have been particularly felt in fields which study complex systems such as the human mind or body using experimental methods. Tribological tests typically aim to investigate properties of or directly compare extremely complex mechanical systems. The global responses of these systems can be altered through a wide variety factors many of which are ignored for practical reasons, cannot be fully controlled, cannot be measured or are neglected to maintain applicability to a real system. Each of these factors has the potential to influence system level results such as forces and wear rates. For this reason results from tribological tests should be expected to show some random scatter and the problems outlined above could be expected in the field of tribology.

There is strong evidence to suggest that typical tribological results show large random variation. For example, current standards for ball-on-flat wear testing [9] and friction testing of plastic sheets [10] give between-laboratory coefficients of variation (COV) of 49% and 18% (static) respectively with in-laboratory values of 23% and 15%. At these levels of variation, if a single test were completed in the same laboratory one would expect the results to be different by more than 50% of the smaller value 23% of the time for wear test (COV=23%). If the tests are performed in different laboratories this rises to

57%. Assuming the COV is accurate, this also means that for ball-on-flat wear testing a total of 21 tests would be required to estimate the mean worn volume, for a lab, with 5% standard error. For more complicated systems results should be expected to contain more variation.

In the light of the problems outlined above, the aim of this study is to assess the state of experimental design in experimental tribology research. In particular the prevalence of simple design fundamentals will be assessed as well as estimates of study power (the chance a study will find an effect if one is present) for the field. This will be pursued through a large scale review of the output from highly regarded journals in the field. The aim is not to critique individual papers or authors, rather just to assess the state of the research and propose improvements.

2. Methodology

2.1. Criteria

The method used to evaluate the quality of each study is given in Table 1. Although other methods have been developed for grading work in tribology [11] these have not been used in favour of more simple and objective measures. Other grading methods also exist for studies in medicine (eg [12]) however, these are also not used as many of the criteria are not relevant to this field.

2.2. Justification for criteria

A brief justification for each of the measures in Table 1 is given below. It should be noted that these criteria have been partly chosen for ease and objectivity of extraction. Thus a study which scores highly on this scale is not necessarily a good study. Factors such as poor instrumentation set-up, unsupported conclusions or logical inconsistencies will not be found. However, a study which performs poorly will be unreliable.

Criteria	Result recorded
Were samples randomised?	yes/no
Were tests randomised?	yes/no
Was any blinding used?	yes/no
Were control tests completed?	yes/no
Were tests repeated?	no/ number
Were repeats on separate batches of materials?	yes/no
Were statistical tests completed?	yes/no
Is full data given?	yes/no
Is data analysis method given?	yes/no
Is data analysis code given?	yes/no
Other recorded information	
Significance	p-value
What was the normalised mean difference (NMD)?	value

Table 1: The review criteria used to assess reliability of research

2.2.1. Randomisation

Test samples should be randomised to remove errors from systematic differences between samples. For example, internal stresses and material properties vary predictably throughout cast billets. Thus, testing a set of samples from the centre of a billet at one contact pressure and comparing them to a set from the edge tested with a different contact pressure confounds the effects of material properties and contact pressure. Such errors can arise from a large number of sources which may be unseen by the experimenter especially when samples are supplied by industrial partners.

The order of tests should be randomised again to avoid confounding intended effects with effects of unmeasured, unseen variables. In tribological testing, systematic errors are particularly likely to be introduced by degradation of testing equipment components or changes in environmental conditions.

2.2.2. Blinding

Ideally the experimenter should be blinded to the type of test/ sample and the expected outcome, both while setting up tests and for any analysis of worn samples. This prevents the experimenter from unconsciously altering the task toward the perceived 'correct' outcome. This is often not possible in full, however, attempts should be made to make techniques as objective or blind as possible, especially when very high resolution techniques are used on relatively large samples, as is the norm in tribology.

2.2.3. Control tests

Control tests should be carried out on identical samples to the main tests with only the variable of interest varying between the tests. These provide a base line against which any changes can be compared. Without a relevant control set of data, any measurement could simply be how a particular material always reacts on a particular test platform. As discussed in the introduction, variation between laboratories is often high so comparing to previously published data is unreliable.

2.2.4. Repeat tests

Repeat tests are essential to ensure that changes observed between groups are not due to random changes or changes correlated with unseen variables. Without repeat tests, no estimation of the variation within a group can be made. An analogy to dice rolling makes this clear. It is easy to imagine rolling two dice and getting a 6 on one and a 1 on the other. However, a good researcher could not conclude that the first dice always rolls higher based on this single sample. Tribological tests are often long and complex with potential for runaway effects. These complexities manifest as a random element in the results, without repeated tests the size of this random element cannot be assessed.

If possible, tests should be repeated on several batches of samples to ensure the observed effects are not due to defects in the original materials or specific properties that are not guaranteed within the specification range. This is espe-

cially important if the control and treatment groups are from different batches of material. For some large scale tests this is prohibitively expensive.

2.2.5. Statistical analysis

Statistical analysis is necessary to indicate how likely the measured effect is to be due to random variation. These tests require repeated or graduated tests to be completed. Although there is much scope for misuse or misinterpretation of these tests, their omission leaves experimenters with no measure of reliability for their outcomes.

2.2.6. Giving raw data, analysis method and analysis code

Pre-publication peer review is not perfect [13]. Showing the full evidence for your research findings by making the raw data and analysis code public allows others to examine findings and check analysis procedures. This also allows for easier collaboration, powerful review articles which collate data from multiple sources, and can help retain methods in research groups when individuals leave. In areas where test samples are very expensive and repeat testing is difficult, data in repositories can be a valuable resource whist sharing code can lead to common analysis norms across research areas, making results directly comparable.

In addition, the p-value associated with any statistical tests were recorded and the normalised mean difference between the control and treatment groups, or the top and bottom value for a parameter sweep type study. The research area was also recorded, however, this was poorly implemented and the results will be ignored apart from where the area was listed as 'not experimental' in which case the paper is removed from analysis.

2.3. Other collected data

In addition to the criteria outlined above the p-value of any statistical tests and the normalised mean difference was collected. A p-value is the result of a statistical test and represents the probability of getting a result at least as extreme as the actual result by chance. Formally is it the chance that the null

hypothesis (typically that there is no effect) is true. This depends on the size of the observed difference, the number of repeats that were run, the amount of variation in the dataset and the assumptions used in the statistical testing. This number is always between 0 and 1, with lower values representing that the result is unlikely to be due to random variation. Historically a p value of 0.05 or lower is taken as *statistically significant* and the minimum threshold at which a positive result can be claimed.

The normalised mean difference is not a standard statistical measure but can be transformed into the *effect size* if the coefficient of variation of the control population is known or can be estimated. The effect size is the difference between the mean of the control group and the treatment group divided by the population standard deviation for the control group. This is typically used as a measure of the importance of a treatment on the measured outcome. The population standard deviation is not typically known but is often estimated as the standard deviation of the control sample. In this study the NMD was collected over the effect size as the effect size cannot be calculated unless the standard deviation is given.

2.4. Selection of works

Works were selected from the top five tribology journals, found by SCImago journal rank (SJR) (retrieved on 20/10/2017) and shown in Table 2. Only works published in 2017 were considered in this study. Works were randomly selected from these journals in proportion to their 2017 output, cropped at 500. The resulting works were then randomly split between 26 assessors. In order to test the inter-rater reliability of our proposed measures, 5% of each assessor's works were also read by another assessor. These works were chosen at random and assessors were blinded to which papers were duplicated. Lists of works were retrieved from Google scholar using Harzing's publish or perish [14], a program for downloading citation data. A python3 program, which is provided in the additional material, was then used to select and allocate works from these lists.

Name	Publisher	SJR	output (2016)
Wear	Elsevier	1.558	301
Tribology International	Elsevier	1.382	672
Tribology Transactions	Taylor & Francis	1.061	199
Tribology Letters	Springer	1.016	71
Journal of Tribology	ASME	0.733	118

Table 2: List of journals which works were extracted from, SJR is as of 20/10/2017

3. Training of assessors

In person training was provided for assessors however after this initial training others joined the project. At a minimum the aims of the project were described over e-mail and the project coordinator (M Watson) was available throughout the process for consultation. All assessors received a formatted answer sheet and a clear description of each criteria with their list of works. Examples of each of these are given in the additional material.

4. Results

Of the 379 unique papers that were assessed, 290 were experimental studies. Where an assessor has recorded that a paper is not an experimental study it is excluded from the following analysis. 'N/A' results are also not included in the following analysis. Basic results for the prevalence of the criteria described above are shown in Figure 1. As shown, prevalence of the majority of the measures is below 10% in this sample, however, control tests (64% n=235), repeated tests (30% n=290) and giving the data analysis method (56% n=264) had a relatively high prevalence. These results are also shown along the diagonal of Table 3.

The above analysis has been repeated for all pairs of criteria, the results of this are shown in Table 3. The top right of Table 3 shows the percentage of works that received a positive response for both criteria out of those that did not receive a N/A rating for either. The number of papers which did not

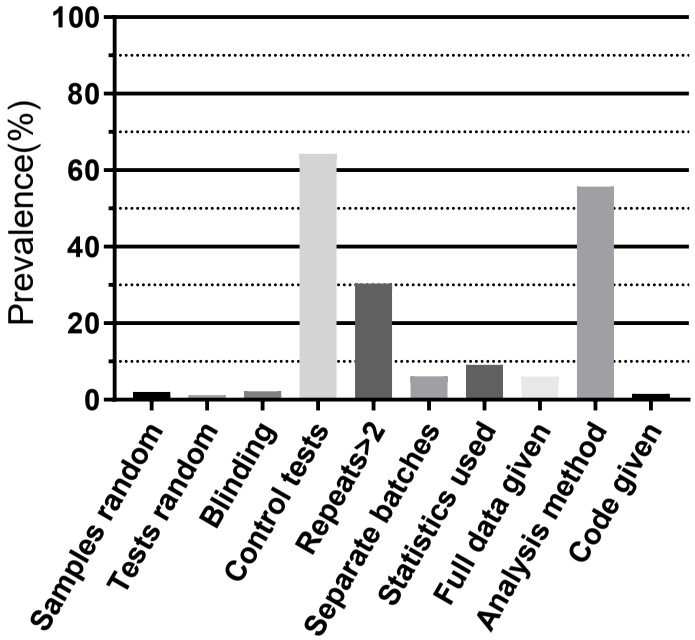

Figure 1: The basic results for prevalence of the measures described above, for each criteria N/A values are ignored

	Samples rand	Tests rand	Blinding	Control tests	n¿2	Separate batches	Statistics	Data given	Analysis method	Analysis code
S R	2 (248)	0 (240)	0 (243)	0.48 (210)	1.6 (248)	0.53 (190)	0.84 (238)	0.41 (242)	1.3 (228)	0 (169)
T R	0	1.2 (256)	0.79 (252)	0.46 (217)	1.2 (256)	0 (196)	0.81 (246)	0 (250)	0.84 (238)	0 (180)
Blind	0	0.29	2.2 (269)	1.3 (226)	1.5 (269)	0 (202)	1.2 (258)	0 (262)	0.82 (245)	0 (182)
Cont	0.0072	0.0072	0.021	64 (235)	22 (235)	5 (179)	8.3 (228)	3.5 (229)	35 (213)	0.61 (163)
n¿2	0.051	0.037	0.046	0.29	30 (290)	4.2 (212)	3.3 (273)	1.1 (283)	20 (264)	0 (193)
Batch	0.059	0	0	0.084	0.11	6.1 (212)	1.4 (207)	1.4 (209)	5.5 (200)	0.69 (144)
Stats	0.08	0.083	0.11	0.12	0.089	0.12	9.2 (273)	0.75 (267)	7.6 (250)	0 (185)
Data	0.056	0	0	0.053	0.03	0.12	0.054	6 (283)	4.2 (260)	0 (192)
A M	0.023	0.015	0.014	0.41	0.3	0.096	0.13	0.075	56 (264)	1.1 (189)
A C	0	0	0	0.0096	0	0.11	0	0	0.023	1.6 (193)
% Yes	1.7	1	2.1	52	30	4.5	8.6	5.9	51	1
% No	84	87	91	29	70	69	86	92	40	66
% N/A	14	12	7.2	19	0	27	5.9	2.4	9	33
Agreement	0.86	1	0.67	0.43	1	1	0.86	0.88	0.57	1
n agree	7	5	6	7	5	4	7	8	7	2

Table 3: Results for each of the categorical criteria studied, values in the top right of the table show the percentage of papers given Yes answers for both of the corresponding criteria where either criteria was graded as not applicable the paper is excluded, the number in brackets is the number of papers not excluded. Values in the bottom left of the table, with a grey background, are the Jackard indices. The proportionate agreement for the repeated measures are also given as well as the number of papers this is based on (excluding papers graded N/A by either assessor)

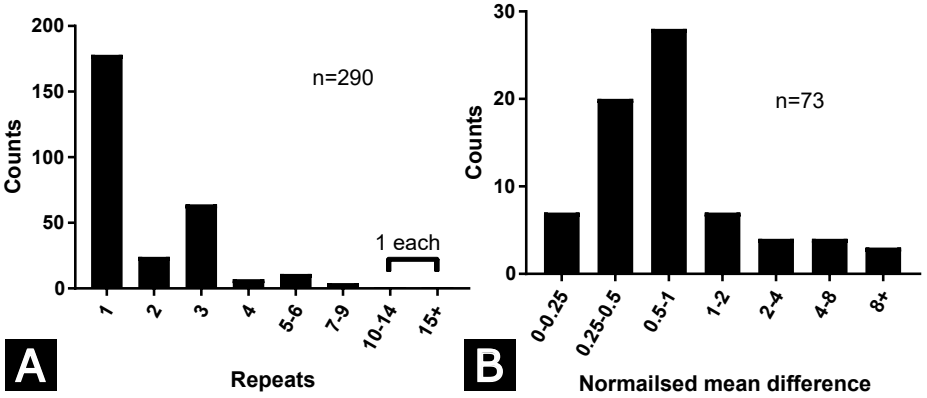

Figure 2: A and B histograms of the sample size and the normalised mean differences recorded

receive a N/A rating for either is given in brackets after the percentage value. The values in the bottom left of this Table are the Jackard indices for each pair. This measure is the intersection between the sets of results divided by the union of the sets, 1 represents a perfect correlation between the measures, 0 indicates no overlap between positive responses.

The proportional agreement between assessors for each of the measures are given in table 3. Whilst statistical analyses for inter-rater reliability exist, these have not been used as they are meaningless when such a large proportion of the results fall into the same category. As shown, these are generally good, however, for the more prevalent criteria, particularly "control tests" and "analysis method given", this measure shows poor agreement between assessors. In addition, for some criteria, particularly "Is analysis code given?", many of the papers were given N/A by one or both of the assessors, leaving a very small sample.

Figures 2A and B show histograms of the number of repeats and the normalised mean difference respectively (NMD). For many of the studies the NMD was not recorded as numerical data were not given and assessors were instructed not to estimate from figures. The median study has a NMD of 0.54 and a sample size of 1.

The collected data can also be analysed by journal, however, χ^2 tests show no suggestive differences between any journal and the mean results for the cate-

gorical data collected at the $p < 0.05$ level, before or after correction for multiple comparisons. A direct comparison between journals could be completed, however, 10 comparisons would be required per measured criteria and the results are likely to be meaningless for a data set of this size.

The full numerical data and the data analysis code used to produce the summary reported above are given in the additional material.

5. Discussion

The results outlined above show that the measures investigated are not well adopted as a whole. Prevalence is below 10% for the investigated factors apart from: control tests, giving details of analysis methods, and repeating tests at least 3 times. For control tests, the agreement between assessors was very poor. This may be due to the definition of control tests used in this study and the variety of studies reviewed. For example it is not always a fair control to compare to bare metal contacts, rather industry standard or best practice should be used as the control, this is not always clear to a reviewer who is not an expert in that field. Likewise, what should be counted as a control test in a parameter sweep is not clear, though this was stated in the instructions to reviewers (included in the additional material).

Analysis methods were frequently given in full detail, however, it should be noted that, due to the lack of statistical tests and repeated experiments, most analysis methods were very simple. For high quality studies with repeats, statistical tests and further characterisation of observed effects the analysis will be more complicated, with options that may invalidate these analyses if chosen incorrectly. In these cases it is imperative to include the full details of the analysis. The analysis code or project files should also be included to ensure that all details are given.

Many researchers in this field are not statistics experts (including the authors of this study) and may not realise the importance of some of the options or default settings. This is particularly important in time series data as including

all of the data will massively decrease the p value of the test but invalidate the analysis as the data are not independent.

Although only 9.2% of studies used statistics, this type of misinterpretation or poor selection of analysis options was observed by some of the reviewers. Of those who presented statistics only 12 works (4.4% n=273) used statistical tests, with others presenting R^2, other summary data or being statistical studies. While most of these presented sound work, some gave inadequate details of the techniques applied [15], failed to correct for large numbers of comparisons [16], misunderstood results [17] or completed tests that were irreverent to the conclusions [18]. Many also ran statistical analysis then presented conclusions that were not supported by the analysis.

Other criteria were very poorly adopted, the lack of randomisation (samples: 2% n=248, tests: 1.2% n=256) observed is particularly imprudent as this is an essentially free step for most experimental studies. In addition to the individual factors, low and zero Jackard indices indicate that there were no perfect studies by these criteria.

It is not possible to estimate the percentage of these research findings that are false as the statistical power is not defined for studies where n=1 and the lack of statistical tests means the threshold for claiming a positive result is unclear. However, as in the introduction, it is possible to investigate how often we would expect such results by random chance, if no actual effects were present.

Figure 3 shows the chance of obtaining such, false positive, results both as a function of the minimum ratio at which a positive result is claimed (r) and the COV in the population. If a COV of 25% is assumed, a result as or more extreme than the median study would be expected in more than 25% of tests on a population in which there are no real effects present.

We can also compare these results to other fields. The reproducibility project in psychology contains relevant data [1], in their sample the median sample size was 52.5 (n=167, data as of 18/7/2018), statistics were used in 93% of studies, in addition many of these studies showed effects though multiple experiments in the same work, so the mean number of experiments presented per study was

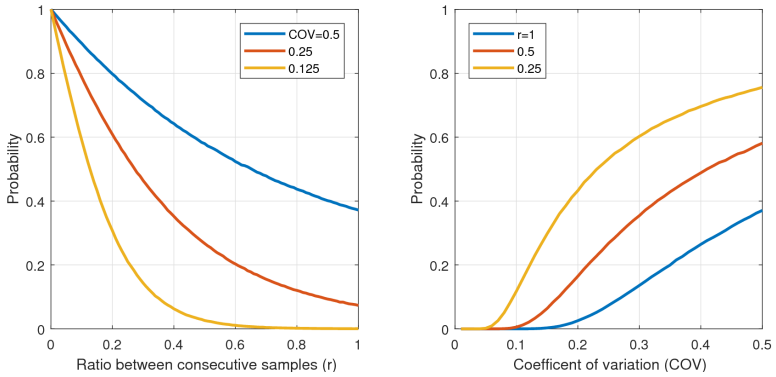

Figure 3: A and B the chance of obtaining a result at least as extreme as the threshold ratio from 2 samples from a normal distribution with the given COV (by simulation)

1.93. Note that this differs from further analysis of worn surface as the samples are independent. Even in this sample the researchers failed to reproduce the results of two thirds of the studies. While this does not mean that two thirds of the studies were incorrect it should give the reader pause, especially considering the comparison with our field.

There is a positive note, if COV of 50% is assumed for the works in this study, the median effect size is close to 2 (NMD/COV). This is a very large effect. In other fields, a large effect is defined as ≥ 0.8 [19]. It is unlikely to be the case that hundreds or even tens of repeat measures are required for reliable results concerning important effects. In addition, many of the hypotheses under investigation are reasonable, and thus relatively likely to be true (high prior probability). This may seem like a subjective measure, and it is, but with a Bayesian view of the statistics this means that many more of the published findings will be true than if more implausible hypotheses were tested [2].

Research is hard, the same flaws we have found in the sample above and many more are present in every experimental study by any of the authors of this work. Our aim is to criticise the standards and norms that plague this research area not individual researchers. We are concerned that many good experimentalists are spending vast amounts of time diligently calibrating, setting

up and running experiments only to have their results confounded by some unseen variable. Or that recommendations are being made to external stakeholders that would not replicate in real life. These problems are damaging to stake holders and the research community but can stem from intelligent people with the best intentions failing to account for the problems described above. Below we outline a way forward that can serve as a reference for designing a high quality study.

6. Recommendations for a high quality study

6.1. Statistical design

The first step in designing a study of a change to a tribological system (such as a change in speed, load, material or coating) is to decide what you are actually interested in. Modern testing machines measure everything possible to give the user more understanding of what is going on. Post test analyses are also often used. This extra information can be useful, but leads to a problem. If the experimenter is biased in any way towards a particular outcome or even just to positive results, measuring many things increases the chance that one of the measured variables will show a desirable change by random variation. This problem is exacerbated with time series data which allows an effectively endless number of comparisons to be made [8]. Having a clear idea of what is of interest before running the experiment narrows the scope to what is important while improving the reliability of any comparisons made.

With a clear variable in mind and an objective, fair test to measure it, rather than proxy measures. The next step is to decide what magnitude of change is of interest [20]. While this is a big problem for other fields, in engineering industrial stakeholders will often have an idea of the cost associated with the change being studied and the value of, for example, an increased service interval. In addition to this, an idea of the typical standard deviation of the population is needed, often this is not known and a reasonable, conservative (larger) value based on experience can be used. The difference between the means divided

by the standard deviation of the control group is the effect size (G. Glass [21], other estimators also exist).

The power of the study should then be chosen [3, 20]. This is the chance of finding an effect of the size that you are interested in if one exists. Formally if α is the chance of a type II error (false negative, missed effect) then the study power is $1-\alpha$. The p-value that will be used to define when a positive result has been identified is also needed. Typically, in the past, this has been 0.05 however, many researchers are calling for this to be lowered to 0.005 to be considered a significant result while results at the 0.05 level should be merely suggestive [5].

With the information from above and the type of statistical test that will be used [4, 8] it is possible to calculate the number of samples needed [20, 22]. This is a relatively complex calculation and will not be included here, many programming packages contain easy to use functions for this purpose, see for example sampsizepwr in matlab, power.t.test in R. For more information on this process [20] provides a clear concise introduction. If more complicated experiments or statistical tests are to be used, simulations of the experiment can be run to find how many repeats are needed. Toolboxes for this purpose exist [23], however, in these cases, it may be worth consulting a statistician before designing the experiment.

While tribology experiments can be expensive and time consuming to this should not be seen as a barrier to good experimental design. While repeated tests can be expensive, there is a trade off between the number of repeated tests per condition and the number of conditions to be investigated. Ultimately this can be thought of reaching several uncertain conclusions or being relatively certain of the most important conclusion.

6.2. Methodological design

With the number of repeated tests decided the rest of the experiment should be designed with the intention of limiting or eliminating any possible confounding factors or bias. Randomisation or systematic variation of the samples and the tests is the simplest way to ensure this for many factors. However, each

experiment will have specific factors that must be addressed, for example milled specimens are often contaminated with some cutting fluid.

Where possible some types of bias can be prevented by blinding the operator to the test parameters or the expected result. While this may seem impossible in tribological research, for many experimental platforms the set-up is identical independent of test condition; loads, speeds and other parameters can be set on a computer just before running. In these cases randomisation and blinding can work well together by simply hiding the test parameters until after the mechanical set up. At this point the operator only needs to enter numbers into a computer and the potential for bias is greatly reduced. If samples are visually identical they can be labelled (A1...,B1...) by a colleague and the meaning of the labels (A-no heat treatment, B-aged...) only revealed after the experiments are completed.

Where post test analysis methods are used, the need for blinding should be seen as directly proportional to the resolution of the analysis machine. Where such small samples of the rubbing surfaces are presented the opportunity for the introduction of concious or accidental selection bias is clear. This again, may seem like a difficult step, however, there are often new researchers who require training on these analysis machines or techniques and have no samples. This can be good means of filling this training gap, fostering better in group collaboration and getting new researchers involved in the publication process. Where possible data or image analysis should also be automated and the code shared.

6.3. Analysis

After (and only after [8]) the experiments have been run the previously determined statistical analyses should be performed [4], again deciding what these should be before getting the results reduces the chance for introducing bias. These can also be blinded, either by hiding categorical names or asking someone else to do the analysis, for further protection against bias. Further tests should not be added to the sample after this point [8].

While it may be tempting to examine other differences that are apparent in the results, the data should not lead the analysis. The problem is that it is not possible to know how many things 'could have been noticed' and therefore the number of comparisons made or that would have been made is difficult to judge [4]. If unforeseen results are of importance it is most conservative to run a second experiment looking for that particular effect.

When a significant effect has been observed this should not be considered as the end of the story. At the very least the data and analysis code should be given to the community so that they may check the calculations and perform meta analyses for important topics. In addition, if possible the result can be validated, generalised or further characterised within the same article, either by scaling or stripping back the experiment to a more fundamental concept. In many other fields it is common to present several experiments in a single work with a common effect between them. This adds both interest and validity to the conclusions.

Lastly, the study should be published. Even if no significant or suggestive effects have been found the presence of a high quality negative study is valuable to other researchers in the same field [6]. While the norm is to publish only positive results, a change with no overall benefit, or even a detriment to performance can appear effective even to well carried out systematic reviews [24, 25, 6].

When publishing, all the data that are collected should be listed, all the experiments that have been done including failed experiments should be described [8]. If data are normalised, omitted or corrected before statistical analysis this should be justified and the results of the statistical analysis without this manipulation should also be presented [8].

7. Conclusions

We intend to change the way researchers think about these experiments, while it is true that samples of mild steel contain less variation than a simi-

lar number of human subjects, there is randomness and contamination present in every sample and procedure. Within a tribological system there is ample opportunity for these random changes to influence macro-scale results.

This study has shown that current research practices in the field do not take this into account, instead these systems are typically researched as if they are purely deterministic and all possible influencing factors are known. Repeated tests and statistical measures are rare while randomisation and blinding are almost unheard of. As such, the potential for confounding factors and researcher bias to influence results remains unmitigated.

We provide a set of simple recommendations for researchers to aim follow. These instructions are not exhaustive or unique [8, 4, 26, 3] but are intended to provide a general approach that is applicable to most of the research in this field. It is not necessary to follow every step for small scale exploratory studies and studies that rely on a strong analysis of a few samples to further mechanistic understanding are still valuable. However, before results of importance or interest are given to stake holders or published a high power study following the recommendations should be completed.

8. Acknowledgements

This work was funded in part by EPSRC grant: EP/R001766/1

References

References

[1] Open Science Collaboration, Estimating the reproducibility of psychological science, Science 349 (6251) (2015) aac4716–aac4716. arXiv:arXiv:1011.1669v3, doi:10.1126/science.aac4716.
URL http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716

[2] J. P. A. Ioannidis, Why most published research findings are false, PLoS Medicine 2 (8) (2005) 0696–0701. arXiv:0208024, doi:10.1371/journal.pmed.0020124.

[3] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, M. R. Munafò, Power failure: why small sample size undermines the reliability of neuroscience, Nature Reviews Neuroscience 14 (5) (2013) 365–376. doi:10.1038/nrn3475.
URL http://www.nature.com/doifinder/10.1038/nrn3475

[4] A. Gelman, E. Loken, The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time, Psychological bulletin 140 (5) (2014) 1272–1280. doi:dx.doi.org/10.1037/a0037714.
URL http://www.stat.columbia.edu/{~}gelman/research/unpublished/p{_}hacking.pdf{%}5Cnhttp://doi.apa.org/getdoi.cfm?doi=10.1037/a0037714

[5] D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, V. E. Johnson, Redefine statistical significance, Nature Human BehaviourarXiv:mky9j, doi:10.1038/s41562-017-0189-z.
URL http://www.nature.com/articles/s41562-017-0189-z

[6] R. Joober, N. Schmitz, L. Annable, P. Boksa, Publication bias: What

are the challenges and can they be overcome?, Journal of Psychiatry and Neuroscience 37 (3) (2012) 149–152. doi:10.1503/jpn.120065.

[7] R. M. Kaplan, V. L. Irvin, Likelihood of null effects of large NHLBI clinical trials has increased over time, PLoS ONE 10 (8) (2015) 1–12. doi:10.1371/journal.pone.0132382.

[8] J. P. Simmons, L. D. Nelson, U. Simonsohn, False-Positive Psychology, Psychological Science 22 (11) (2011) 1359–1366. arXiv:2021, doi:10.1177/0956797611417632.
URL http://journals.sagepub.com/doi/10.1177/0956797611417632

[9] ASTM, ASTM G133 - 05 Standard Test Method for Linearly Reciprocating Ball-on-Flat Sliding Wear, ASTM International, West Conshohocken, PA, 1995, 2016. doi:10.1520/G0133-05R16.
URL http://www.astm.org/cgi-bin/resolver.cgi?G133-05(2016)

[10] BSI, BS EN ISO 8295:2004 Plastics — Film and sheeting — Determination of the coefficients of friction, The British Standards Institution, 2004.

[11] M. Harmon, R. Lewis, Review of top of rail friction modifier tribology, Tribology - Materials, Surfaces and Interfaces 10 (3) (2016) 150–162. doi:10.1080/17515831.2016.1216265.
URL http://dx.doi.org/10.1080/17515831.2016.1216265

[12] V. W. Berger, S. Y. Alperson, A general framework for the evaluation of clinical trial quality., Reviews on recent clinical trials 4 (2) (2009) 79–88. doi:10.2174/157488709788186021.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2694951{&}tool=pmcentrez{&}rendertype=abstract

[13] Springer, SCIgen-generated papers in conference proceedings (2014).
URL https://www.springer.com/gp/about-springer/media/statements/scigen-generated-papers-in-conference-proceedings/26276

[14] A. Harzing, Publish or Perish (2007).
URL http://www.harzing.com/pop.htm

[15] L. Zheng, J. Peng, J. Zheng, D. Liu, Z. Zhou, Surface properties of eroded human primary and permanent enamel and the possible remineralization influence of CPP-ACP, Wear 376-377 (2017) 251–258. doi:10.1016/j.wear.2017.01.055.
URL http://dx.doi.org/10.1016/j.wear.2017.01.055

[16] A. Borjali, J. Langhorn, K. Monson, B. Raeymaekers, Using a patterned microtexture to reduce polyethylene wear in metal-on-polyethylene prosthetic bearing couples, Wear 392-393 (July) (2017) 77–83. doi:10.1016/j.wear.2017.09.014.
URL http://dx.doi.org/10.1016/j.wear.2017.09.014

[17] R. Namdeo, S. Tiwari, S. Manepatil, Optimization of High Stress Abrasive Wear of Polymer Blend Ethylene and Vinyl Acetate Copolymer/HDPE/MA-g-PE/OMMT Nanocomposites, Journal of Tribology 139 (2) (2017) 021610. doi:10.1115/1.4034696.
URL http://tribology.asmedigitalcollection.asme.org/article.aspx?doi=10.1115/1.4034696

[18] M. Buciumeanu, J. R. Queiroz, A. E. Martinelli, F. S. Silva, B. Henriques, The effect of surface treatment on the friction and wear behavior of dental Y-TZP ceramic against human enamel, Tribology International 116 (July) (2017) 192–198. doi:10.1016/j.triboint.2017.07.016.
URL http://dx.doi.org/10.1016/j.triboint.2017.07.016

[19] J. Cohen, A Power Primer, Psychol Bull 112 (July) (1992) 155–159. arXiv:arXiv:1011.1669v3, doi:10.1038/141613a0.

[20] S. R. Jones, An introduction to power and sample size estimation, Emergency Medicine Journal 20 (5) (2003) 453–458. doi:10.1136/emj.20.5.453.
URL http://emj.bmj.com/cgi/doi/10.1136/emj.20.5.453

[21] G. V. Glass, B. McGaw, M. L. Smith, Measuring study findings, in: Meta-analysis in social research, SAGE Publications, Ltd, Beverly Hills, CA, 1981, Ch. 5, p. 126.

[22] S. M. Govani, P. D. Higgins, How to read a clinical trial paper: A lesson in basic trial statistics, Gastroenterology and Hepatology 8 (4) (2012) 241–248.

[23] P. Green, C. J. Macleod, SIMR: An R package for power analysis of generalized linear mixed models by simulation, Methods in Ecology and Evolution 7 (4) (2016) 493–498. doi:10.1111/2041-210X.12504.

[24] B. Hart, A. Lundh, L. Bero, Effect of reporting bias on meta-analyses of drug trials: Reanalysis of meta-analyses, BMJ (Online) 344 (7838) (2012) 1–11. doi:10.1136/bmj.d7202.

[25] J. P. Ioannidis, Why most discovered true associations are inflated, Epidemiology 19 (5) (2008) 640–648. doi:10.1097/EDE.0b013e31818131e7.

[26] D. Moher, K. F. Schulz, D. G. Altman, The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials, The Lancet 357 (9263) (2001) 1191–1194. doi:10.1016/S0140-6736(00)04337-3.