The Role of Hypotheses in Biomechanical Research

Darrell P. Rowbottom (Lingnan) & R. McNeill Alexander (Leeds)

DarrellRowbottom@ln.edu.hk

R.M.Alexander@leeds.ac.uk

Abstract

This paper investigates whether there is a discrepancy between the stated and actual aims in biomechanical research, particularly with respect to hypothesis testing. We present an analysis of one hundred papers recently published in *The Journal of Experimental Biology* and *Journal of Biomechanics*, and examine the prevalence of papers which (a) have hypothesis testing as a stated aim, (b) contain hypothesis testing claims that appear to be purely presentational (i.e. which seem not to have influenced the actual study), and (c) have exploration as a stated aim. We found that whereas no papers had exploration as a stated aim, 58% of papers had hypothesis testing as a stated aim. We had strong suspicions, at the bare minimum, that presentational hypotheses were present in 31% of the papers in this latter group.

1. Introduction

In this paper, we attempt to determine to what extent there is a discrepancy between the stated and actual aims of research in biomechanics (i.e. the application of the theories and methods of mechanics to the study of living things). In particular, we are

interested in determining how often work is presented as testing a hypothesis (or hypotheses), when the intent was not to test said hypothesis (or hypotheses).

We are interested in 'hypothesis testing talk' because anecdotal evidence suggests that research in the biological sciences is more likely to be funded, and even to be published, if it involves – or, rather, appears to involve – hypothesis testing. For example, a highly-regarded biomechanist wrote in an e-mail to Alexander that "BBSRC [The Biotechnology and Biological Sciences Research Council] are focused on funding hypothesis driven research. As most of my work has been funded by the BBSRC, it is largely hypothesis driven." (The individual in question was basing this claim on his discussions with a named colleague who has served on the relevant BBSRC committee.) However, we also believe that natural scientists sometimes go on what might be called 'fishing trips'; identifying and examining areas where looking carefully might disclose something interesting. We would therefore expect such 'fishing trips' to happen in several biological and biomedical contexts, whether or not they are presented as such.

In fact, we have evidence that such 'fishing trips' do occur in biomechanics from a survey of eleven well-regarded biomechanists with a range of experience, which we conducted before embarking on our analysis of published papers. We asked each respondent to state the actual aims behind the work that led to their three best papers, and were told that in five cases (15% of the total) this was to explore a topic rather than to test a hypothesis, collect specific data, or answer a specific question. One respondent even suggested that:

> [T]he balance is skewed by you asking for my best papers. If you asked for my last three or ten papers then, especially with some work undertaken by research students there would be more of the 'go and measure these parameters' … and then examining the data looking for a pattern. A cynic might call that a fishing expedition...

Another explained how his exploration bore fruit:

> Although [a colleague] and I started this investigation more to explore the topic and find out more about it, I remember that suddenly, one day I realized that by combining [the data with known muscle efficiencies] the experimental minimum of metabolic cost could be predicted.

And finally, a third recounted the following tale:

> [A senior colleague] invited me to attend [a particular conference]. I could not think of anything I could say [about the topic] and [he] suggested I collected data on … The aim of enabling me to attend the meeting was achieved! … My outlook had changed by the time we came to write [the paper]. We now had a specific question to formulate and answer. I could not have posed the question prior to collecting the data.[1]

[1] In the interests of balance, we should add that a fourth respondent, who classified two of his best papers as hypothesis testing, wrote: "I have surprised myself a little [in answering your questionnaire]! I have always imagined that my science had a very large measure of floundering in the dark, but an analysis of three of my favourite papers suggests otherwise. But perhaps it is because the work was so sharply focused from the start that it yielded the more satisfactory publications." Note, however, that what scientists consider to be their best work may be influenced by the citation count. And citation count may be affected by whether the paper claims to test a hypothesis. These matters are worthy of empirical study, but are beyond the scope of the present paper.

It should be noted from the start that it is normally rather easy to present research as testing a hypothesis (or hypotheses) even when it was not really designed for that purpose. Imagine that a social scientist conducts a study on how salary varies with respect to educational achievement in the UK, and that she relies on questionnaires in order to collect data. But now imagine she notices something that she did not expect, namely that quality of handwriting correlates inversely with educational achievement. When she comes to presenting her research, she might easily suggest that it was designed to test a hypothesis (or hypotheses) concerning handwriting; and this might be a good strategy to improve probability of publication. (Indeed, she might even suggest that she gathered qualitative data to increase the amount of handwriting on each questionnaire.)[2] Alas, however, her readers would be misled about how she worked if they took her paper at face value.

Such concerns have been raised before, of course. Popper (1983, 47–48), for instance, once bemoaned the inductive style that he saw as prevalent in biological research:

> Nowhere is the power of the inductivist tradition as conspicuous as in what I have called '*the inductive style*' – a certain manner of reporting one's researches which is still the traditional way of writing in a number of biological journals, although by now it has almost disappeared from the journals of physics and chemistry.

[2] Similar examples are easy to construct, and are not peculiar to social science. Here's a medical one: a researcher wishes to investigate how effective a drug is in treating a specific condition, but is told by the drug manufacturer, subsequent to the completion of her study, that all the pills she was provided with were defective. In an attempt to prevent all her hard work going to waste, she might present the study as testing a hypothesis about the placebo effect.

> The basic idea which inspires the inductive style is this: we must keep carefully to our actual observations, and must beware of theorizing; for this may make us acquire theoretical prejudices which may easily bias or taint our observations if we are not *very* careful...
>
> [N]o doubt those trained to write in this way are unaware that this laudable and apparently safe idea is the mistaken result of a prejudice – worse still, of a philosophical prejudice – and of a mistaken theory of objectivity... [W]e cannot avoid or suppress our theories, or prevent them from influencing our observations; yet we can try to recognize them as hypotheses and formulate them explicitly, so that they may be criticized.[3]

As a concrete example, Popper gave Fleming's (1929) classic paper in which the discovery of penicillin is reported. He objected that Fleming's presentation makes it seem as if the discovery was entirely unexpected, although theories concerning antibodies had been common currency for almost a century beforehand:

> Fleming's discovery was not really accidental: it was the work of a great discoverer who knew very well what he was doing, and what was worth describing: and though it was an accident that the mould whose antibiotic properties he had observed turned out to be non-toxic, the existence of substances of this kind had been expected, and hoped for, for a long time. This expectation motivated the work both of Fleming and [subsequently] of Florey's team. (Popper 1983, 49)

[3] It should be noted that although this discussion was first published in 1983, Popper completed it much earlier; and by 1962 at the latest. See the editor's (W. W. Bartley's) foreword in Popper 1983.

Popper was writing around fifty years ago, however. (See n.3; Medawar 1963, which is better known, came slightly later!) So might the situation now be rather different? Perhaps Popper's views – (1) that scientists should state their assumptions clearly and explicitly, so that they may be easily criticized; and (2) that genuine scientific activity involves only conjecture and refutation – have even been responsible for some changes. Is a (broadly) hypothethico-deductive style where research is made out to involve more predictions and tests than it actually does, and which is just as misleading as its inductive forerunner, the new usurper?[4]

Why is this investigation worthwhile? As we will later illustrate, we take its results to have potential epistemic, and not just stylistic or sociological, significance. In particular, many philosophers and scientists have thought that theories cannot be confirmed (or even corroborated) by old data that they merely *accommodate*. The nub of the idea is that it is one thing to generate a theory that is consistent with the data one happens to have, or to 'data fit'; it is quite another then to generate novel predictions using that theory, and to put it to the test. According to Popper (1959), for instance, how we arrive at our theories is a matter for psychological study only; what we do with them, when we have them, is the important epistemic part. And if we were to be given a false impression about what had been done with a theory, then we might also be given a false impression about how we should regard it. An historical example might help to illustrate the issue. Leverrier's prediction of the existence of Neptune – see Hanson (1962) – was heralded as a great success for Newtonian mechanics. But would it have been as great a success if Neptune had been spotted first, and Leverrier

[4] We think that Popper was correct about (1), but that his position on (2) was too strong. See Rowbottom (2011a), for example, on how both dogmatic and critical scientists may be necessary for the best possible science.

had subsequently shown that Newtonian mechanics could accommodate, or was consistent with, its existence? (It may help to think about how astrology managed to accommodate Neptune's existence with relative ease.) And how about if Neptune had been known about for many years before Newton was born?[5]

Before we continue, we should also say something about why we suspect that examining biomechanics is an interesting way of shedding light on this issue. The answer is simple. Since biomechanists rely on physical theory to a large extent in conducting their work – indeed, several have backgrounds in physics or engineering – we might expect their work to be similar to that of other physicists and engineers, i.e. not especially aimed at testing specific hypotheses (whether or not it ought to be).[6] So to the extent that it is dissimilar, we might expect that this is as a result of involvement in, and interaction with, biological scientists. Ultimately, however, this study is interesting in its own right in so far as it provides evidence of a peculiar form of emphasis (or even misrepresentation) in biomechanical publications.

2. Method

For our investigation we selected two journals that we consider to be among the most important for the subject of biomechanics: *Journal of Biomechanics* and *The Journal*

[5] This is a considerably more complicated example than it may first seem; it may also matter, for example, at what point Uranus' orbit was known. One issue is whether the prediction must be *temporally* novel, or need merely have been a prediction that was never made before – i.e. which was never previously derived from any physical theory and appropriate background assumptions. If the latter view is correct, then one would do better to consider two theories that predict some of the same things. So we might compare Newtonian mechanics with relativity, and ask whether our old data concerning the accuracy of relativity when the gamma factor is approximately zero serves to confirm/corroborate relativity.

[6] We do not provide support for this contention that the work of physicists and engineers is not generally aimed at testing specific hypotheses, because another study of similar nature to that presented here would be required.

of Experimental Biology. *Journal of Biomechanics* publishes only biomechanical papers, most of them on or relevant to human biomechanics. Typical authors work in engineering or medical science laboratories. *The Journal of Experimental Biology* publishes papers not only on biomechanics, but also over a much wider field. Most of its papers are about non-human animals, but there are also some on human biomechanics. Typical authors work in biology laboratories. For decisions as to which papers in *The Journal of Experimental Biology* should be regarded as biomechanical in subject matter we relied on the judgement of Alexander, who has fifty years experience in biomechanical research and has written several textbooks in the field.

We each classified the papers independently, using three categories for stated aims:

(1) *Hypothesis Testing* (H)

(2) *Exploratory* (E)

(3) *Other* (O)

With respect to the classification of H, we decided that the philosophical question of what genuinely counts as a test, or an attempt to test, is beside the point. We are interested in whether scientists attempt to present their work as testing a hypothesis (or hypotheses), and therefore only in what they *perceive* to fit the bill. This mitigates in favour of allowing both 'hypothesis' and 'testing' a broad scope. As such, we agreed that professed attempts to 'verify', 'validate', or 'confirm' all counted as professed attempts to test. Furthermore, we agreed that 'a hypothesis' might be a prediction, a theory, or even a model (e.g. a computer model). Finally, uses of the verb 'hypothesize' were typically understood to imply that the paper sought, in part if

not in whole, to test something. Exceptions were only made when 'hypothesize' was used to state a working assumption, e.g. a theory that was taken for granted in the execution of the research.

We also decided that we would classify a paper as H if there was *any* mention of hypothesis testing (using the broad understanding explained above) as an aim, even as subsidiary and among many. But we did not, however, give such a classification if the paper claimed only that some collected data bore on a hypothesis (e.g. by 'confirming', 'supporting', 'corroborating', or 'disconfirming' it). Recognizing that there would still be differences in classifications, we also agreed that we would discuss these in order to resolve them, when they occurred, and leave open the option of introducing a 'disagreement' category for cases in which we were unable to reach a considered consensus.

Papers put into the H category would be subsequently examined, again independently in the first instance, in order to see in how many instances we would each classify the stated aim of hypothesis testing as merely *presentational* (P), rather than actual. In short, we were looking for cases in which the stated aim (of H) did not match the actual aim. We agreed that a P classification would be given if *any one hypothesis testing statement* in the paper appeared to be presentational.

However, it soon transpired, when we compared and discussed our results, that there were several cases in which we thought that it was appropriate to register a serious *suspicion* (S) that a hypothesis was presentational, without giving a full P classification. (The difference between the two classifications is in estimated evidence

levels. We classified a paper as P if we thought there was strong evidence of the presence of at least one presentational hypothesis. We used an S classification if we thought there was some evidence, but not strong evidence, of such a hypothesis. The difference could be expressed numerically, by an appropriate function, e.g. a probability function, ranging from zero to one. Roughly, a P classification corresponds to a value above 0.8. An S classification corresponds to a value in the interval [0.5, 0.8].) We will discuss some specific examples in the next section.

We would also like to emphasize that we erred on the side of caution in classifying papers as involving presentational hypotheses (P), so as to avoid false positives (at the risk of an increase in false negatives). If either of us had doubts about whether a P classification was warranted after the discussion phase of our data analysis, in any particular case, then we only registered a suspicion (S) at most. (Thus if one of us scored a paper as P, and the other scored it as S, it would not have a final P classification.) Furthermore, a suspicion must have been agreed as reasonable by us both, in order for an S classification to be awarded. (Hence if one of us scored a paper as S, and the other scored it as neither S nor P, then it would not have a final S classification.)

Clearly the judgements required were difficult, and we were error prone in making them. We could not know the minds of the authors! However, we could isolate cases where (1) it seemed abundantly clear before the research reported in the paper was performed, often by the author's own admission, that one of the stated hypotheses was

true (or false)[7], and/or (2) the alleged intent to test a hypothesis appears to have been incidental, at best, to the research as a whole. We should also like to make it clear that such cases may not involve any form of significant scientific misconduct. For example, research findings can falsify a hypothesis even if they were not intended to do so; so to misrepresent an experiment as an intentional test of *T* need not mislead anyone about whether *T* was falsified. Rather, it would only make out that an accidental or incidental finding was actively sought.

Nonetheless, in some cases such misrepresentations might mislead about to what extent a hypothesis has been confirmed/corroborated by an experiment, at least on some accounts of scientific confirmation. For example, it has often been held that the real test of a theory or hypothesis is whether it yields a successful risky prediction (i.e. a prediction of something previously unexpected), rather than merely accommodates known data. As Musgrave (1974) notes, versions of this view have been held by many philosophers and scientists, such as Descartes, Leibniz, Whewell, and Duhem. There are also others who do not appear on his list, who at least thought that prediction is significantly different from accommodation in epistemic import. For Boyle (cf. Stewart 1991, 119), for example, one quality of an excellent hypothesis is:

> That it enable a skilfull Naturalist to Foretell Future *Phaenomena*, by their Congruity or Incongruity to it: and especially the Events of such Expts as are aptly devisd to Examine it; as Things that ought or ought not to be Consequent to it.

[7] Models cannot, of course, be true or false; in these cases, 'good' or 'bad' (or 'fit for purpose' or 'not fit for purpose') would be more appropriate. Similarly, one might read 'accurate' or 'inaccurate' for predictions.

As Musgrave (1974, 5) notes, "Popper occasionally suggests that it is a psychological affair: for a test to be severe the experimenter who performs it must be sincerely trying to overthrow the theory tested." But even if it is not quite such a psychological affair, the background knowledge at the time the experiment is performed (either of the individual, or more plausibly in the field) is arguably crucial.[8] This is suggested by Popper's (1959, 1983) measure of corroboration, the workhorse of which is P(*e*, *h* & *b*) - P(*e*, *b*), where *e* is a report of an experiment, *h* is a hypothesis, and *b* is background knowledge at the time of the experiment. A classic example of excellent corroboration (where P(*e*, *h* & *b*)>>P(*e*, *b*)) was the discovery of the Poisson (or Arago) bright spot, which was predicted on the basis of Fresnel's wave theory of light but was not expected to be found (especially by Poisson), as Rowbottom (2008a; 2011b, 45–48) explains. Rightly or wrongly, and this is a matter we will say no more on here, a theory of light which had merely accounted for what was already expected would not have been so impressive. For further recent discussion on the relative value of prediction and accommodation, see Barnes 2005, Hudson 2007, Harker 2008, and Rowbottom 2008b, 2011b.

3. Results

In total, we examined one hundred papers. We took fifty papers from volume 41 (2 & 6) of *Journal of Biomechanics* (2008), and fifty papers on biomechanics from volume 210 (21–24) and volume 211 (1 & 3–7) of *The Journal of Experimental Biology* (2007 & 2008). As noted above, relevant papers from the latter were selected by Alexander.

[8] See Rowbottom (2011b, 94–95) on the question of whether sincerity matters. The main argument that it does, in summary, is that sincerely searching may cause one to look in places that one otherwise might not. (And true observation statements may nonetheless be *selective*.) This gels with the idea that we are prone to confirmation bias unless we design appropriate experiments.

During our deliberations, we encountered two unanticipated problems in a small number of cases. The first of these concerned how we should classify a paper which stated that its aim was to 'test whether' something is the case (rather than 'test the hypothesis' that it is, or is not, the case). Almbro and Kullberg (2008), for example, state that they "tested whether the flight performance of an insect … is affected by variation in body mass due to feeding." This was a particularly interesting case, for us, because we were both confident that the paper would be given a P classification were it to be given an H classification (and one of us classified it as H & P in the first instance, but the other classified it as O). This is because the authors already knew by their own admission elsewhere in the paper (a) that the change in body mass of the insects in question after feeding could be as great as 50%, and (b) that previous studies had shown that far smaller increases in mass in the relevant area of the body resulted in decreased flight performance. (It is also pretty obvious by analogy, even to a layman, that a 50% increase in the mass of an aircraft after loading, roughly concentrated in the belly area, would have an effect on its flight performance!) So it appears that they really intended to *quantify* the (negative) change in flight performance due to increased body mass after feeding, rather than to test the hypothesis that any change (positive or negative) occurs. (And they were interested in the extent of the change, we think, with respect to the 'fitness' of the organism in question when it comes to escaping predators.) So we both agreed that Almbro and Kullberg (2008) did not intend to test what they say they tested, but not that they presented their work as testing a *hypothesis*. Rowbottom thought that they were clearly making out that that they were testing *either* the hypothesis that "flight performance... is affected by variation in body mass due to feeding" or its negation.

He also thought that it was reasonably clear that they purported to be testing the former (by the context). Alexander, on the other hand, maintained that there is a significant difference between 'testing whether *p*' and 'testing the hypothesis that *p*'. Eventually we decided that we should not classify that paper as H, in line with our charitable policy of favouring false negatives (especially for S and P) over false positives. We mention this issue here, however, because if Rowbottom is correct then presentational hypotheses are more prevalent than our results suggest.

The second problem, which was related, occurred when authors said that they wished to 'determine whether', rather than 'test whether', something is the case. Rowbottom originally gave H verdicts for some of these papers, whereas Alexander gave only O verdicts. After discussion, we agreed that O verdicts were preferable because there was not even any explicit mention of testing. Again, this was in line with our policy of favouring false negatives.

A summary of our overall results follows:

Table 1 – Final Classification of 100 Papers from *The Journal of Experimental Biology* and *Journal of Biomechanics*

	H (Hypothesis Testing)	E (Exploratory)	S (Suspected Presentational Hypotheses)	P (Presentational Hypotheses)
The Journal of Experimental Biology	**27**	**0**	**4**	**3**
Journal of Biomechanics	**31**	**0**	**8**	**3**

Table 2 – Percentages of H, S & P Papers

	H	S	P
The Journal of Experimental Biology	**54%**	**8%**	**6%**
Journal of Biomechanics	**62%**	**16%**	**6%**

Overall	**58%**	**12%**	**6%**

Table 3 – Percentages of H Papers Classified as S, P, or S or P

	S	**P**	**S or P**
The Journal of Experimental Biology	**15%**	**11%**	**26%**
Journal of Biomechanics	**26%**	**10%**	**35%**
Overall	**21%**	**10%**	**31%**

We will postpone our concluding discussion of the significance of these findings until we have discussed some examples of papers classified as S and P.

4. Examples of Presentational and Suspected Presentational Hypotheses

The papers in which we found presentational hypotheses are as follows: Astley and Jayne 2007, Clark and Summers 2007, Bates et al 2008, Nowlan et al 2008, Sigal et al 2008, and van der Merwe et al 2008. We discuss three of these in greater depth below.

First, Astley and Jayne (2007) supposedly hypothesised that "snakes on cylindrical branches would use a form of concertina locomotion" and "expected that the mean

velocity would decrease when moving up steep inclines" (which we take, in context, to be a second hypothesis). We regard both hypotheses as presentational because it was already well-known that snakes on branches use concertina locomotion and because it would be astonishing if crawling were not slower on steep inclines. After explaining that there are three types of concertina locomotion in a reference text intended for advanced students, for example, Edwards (1985, 167) then states that one: "type of surface concertina locomotion is used by all limbless tetrapods on flat substrates on which projections are either too small or too sparse to act as good pivotal points. It is also used by snakes in traversing thin structures, such as branches and even telephone wires, where pegs are not available for lateral undulation."

Second, Clark and Summers (2007) studied hagfishes, primitive chordates which have tooth plates but no true jaws. They suspected that these tooth plates may be inferior to jaws, and considered why this might be. They reported observations on more than one species of hagfish and advanced the hypothesis that the two closely related species are similar (in relevant respects). However, this is a common assumption in biology, and we regard its formal expression here as presentational. (We only know they are closely related species, arguably, because they are similar!)

Third and finally, Bates et al (2008) 'hypothesized' that a part of the echolocation call of a species of bat would be vulnerable to interference from conspecifics using nearby frequencies, and that bats would therefore have a "jamming avoidance response" (or "JAR"). But the problem of interference between bats using similar frequencies is well known, and it is inevitable that an echolocation system will be jammed by nearby frequencies from another source. If we add that bats function in an unimpaired fashion

in large groups, it becomes obvious that they must have a JAR. Bates et al (2008) claim that: "Taken together, the existing observations of changes in broadcast frequency by bats flying in groups or responding to playback in the field do not provide conclusive evidence for a JAR in bats." However, their standard for "conclusive evidence" seems entirely unrealistic, especially since they admit elsewhere in the piece that:

> [V]ideos of swarming bats indicate that they have little problem orienting and capturing prey in the presence of many other echolocating bats (Simmons et al., 2001). Animals that emit their own orienting signals could adapt by changing the frequencies of their signals in the presence of interference in order to avoid masking or 'jamming'...
>
> Some investigators have reported greater differences in emitted frequency between two bats of the same species flying in close proximity than between two randomly selected single bats of the same species... Bats flying in groups have been observed to change the duration of their pulses or their inter-pulse intervals... as well as the frequencies of their broadcasts. (Bates et al 2008)

Furthermore, Griffin et al (1963) had long ago, in a paper to which Bates et al (2008) refer, noted that: "Often echolocation is complicated by orientation sounds from other bats nearby."

We also strongly suspected the presence of (one or more) presentational hypotheses in: Ellerby and Askew 2007, Estrella and Masero 2007, Hedrick et al 2007, Anders et

al 2008, Chang and Ulrich 2008, Ford et al 2008, Jenkyn et al 2008, Moazen et al 2008, Parsons et al 2008, Siegmund et al 2008, Tan et al 2008, and Verhulp et al 2008. Below, we discuss four of these papers:

First, Anders et al (2008) investigated muscle activity in the human trunk when the body was tilted. They noted that trunk muscles have been classified as stabilising or mobilising and hypothesised that that these two classes of muscles would employ different strategies. Their results ran counter to the hypothesis. This would seem at first sight to be a classic test of a hypothesis, but the predictions are so specific – that the EMG-force relationship would have "a linear characteristic for local muscles and non-linear curves for the global muscles" (Anders et al 2008) – that Alexander suspected that the hypothesis may only have been formulated after the results were known, in which case it should be classed as presentational. Alexander was not aware of any reason why one would expect such a difference.

Second, Estrella and Masero (2007) studied distal rhynchokinesis, a mechanism that enables long-billed wading birds to open the tip of their bills independently of the other parts. They "predicted that the protraction of the bill tip [the mechanism that drives rhynchokinesis] during the strike and transport phases would be greater with larger sized prey." (ibid.) This seems so obvious, and so trivial in relation to the main aim of the paper (which is to show that the birds use distal rhynchokinesis to feed on items suspended in the water), that we have a strong suspicion that it was presentational. The only doubt, expressed by Rowbottom, arises because it is possible to open the bill in two ways (i.e. also along its length). However, there are clear and reasonably obvious advantages to using distal rhynchokinesis; it does not, for

instance, allow so much water to enter the bill from the sides (which would reduce the efficiency of suction feeding).

Third, Chang and Ulrich (2008) discussed a disease that severely impairs human gait. They presented hypotheses about the effects on patients' gaits of cords that restrict lateral sway. Specifically, the authors 'hypothesized' that the cords reduce sway and its variability, decrease co-activation of muscles and reduce energy expenditure. These hypotheses seem presentational, given that previous tests on healthy subjects had shown that such cords reduced sway variability and metabolic costs. (Chang and Ulrich cite Donelan et al 2004 on this point, so were plainly aware of it.) The authors' intent does not seem to have been to test the hypotheses. Rather they appear to have wanted to find out how and to what extent the patients' gait was improved.

Fourth and finally, Siegmund et al (2008) discussed the effects of cross-links on the failure properties of mineralised collagen fibrils. They "hypothesise[d] that predicted stress-strain curves under conditions with no collagen cross-links, with only enzymatic cross-links ... or with additional non-enzymatic cross-linking ... would exhibit significantly different characteristics." This is remarkably vague, but it is well-known that cross links increase the strengths and elastic moduli of polymers. (An analogy with bamboo scaffolding might even suffice for seeing this.) Their real aim seems to have been to measure these changes, rather than to check whether they really occur.

We can therefore see, in each of the seven previous examples, that there were two main reasons for which papers were eventually classified as P or S. In a small number

of cases, the data did not appear to bear on one or more of the hypotheses that it was the stated intention to test. In almost all of these cases we registered only a suspicion (S), because we were concerned that the authors may have taken their work to have relevance to all the stated hypotheses (although it did not in our opinion). In the other cases, which were more common, one of the stated aims of the paper was to test a hypothesis that was widely accepted to be true or false – so widely, indeed, that a paper which stated *simply* that it would test the hypothesis would never have been considered interesting enough to be published unless it ran counter to the expectation – when it appears that the actual point of the exercise was to gather data in order to understand some phenomenon (or phenomena) better. A sub-set of these cases were hypotheses that we thought to be *post hoc*; typically, in such cases, we registered only a suspicion.

Recall that we erred on the side of caution when it came to judging whether a stated hypothesis was merely a working assumption. In their discussion of the echolocation behaviour of bats approaching prey or a landing site, for example, Melcón et al (2007) stated the rather vague "working hypothesis" that "the difficulty of the echolocation task is reflected in the approach behaviour." The next two sentences are more explicit and can be read as part of the hypothesis but merely postulate, more or less, that bats will behave as observed in previous studies. If the "working hypothesis" was intended as a formal hypothesis, it is presentational. However, we gave the authors the benefit of the doubt and did not register the paper as H.

5. Discussion

Our most striking findings, over the one hundred papers, are as follows. First, no papers had exploration as a stated aim.[9] Second, 58% of papers had hypothesis testing as a stated aim. Third, out of those papers which had hypothesis testing as a stated aim, approximately one third (31%) were strongly suspected, at least, of containing some purely presentational hypotheses. We will discuss these findings in turn.

The first finding is perhaps the most remarkable, because it indicates a strong bias in biomechanics against presenting work as the result of exploratory activity. That is, given our evidence from survey work, mentioned in the introduction, that biomechanists will privately admit that purely exploratory activity does lead to published work. Obviously there are two primary means by which this might happen: on the one hand, authors may intentionally avoid any reference to exploration as an aim of their studies; and on the other, mentions of exploration may be filtered out by the refereeing process.

Recall that exploration appears to be directly opposed to hypothesis testing in the following sense: to look somewhere in the hope of finding something interesting does not require any hypotheses about what one will find, and does not involve the explicit intent to test of any hypotheses whatsoever (other, perhaps, than trivial non-scientific hypotheses of the form "We will find something interesting if we look at *x*"). Thus the evidence that biomechanists will avoid presenting their work as exploratory is consistent with (and arguably corroborates, in conjunction with the other findings) the view that hypothesis-testing work is up-valued. In short, to fail to present one's work as hypothesis-testing is not to rule out that it involved testing some hypothesis. But to

[9] The only paper that came close to being classified as E, and then only by Alexander, was Marshall et al 2008. One of their stated aims was to "investigate feeding behaviour in bearded seals to determine the range of their behavioural repertoire for capturing prey."

present one's work as exploratory *is* to be explicit that the work was not aimed at testing some hypothesis (or hypotheses). (Needless to say, it is possible that exploratory work could be frowned upon independently of favouring hypothesis-testing work. But admitting an exploratory strategy would not appear to be precluded by adopting the inductive style written of by Popper and Medawar, for example.)

The second finding is also noteworthy, in so far as it is not terribly plausible that well over half the work (worthy of reporting) that goes on in biomechanics in genuinely aimed at testing hypotheses. This may be highlighted by our finding, in a pilot study, that roughly 25% of forty papers in the final sample had data collection as a stated aim.[10] This would leave only 17% of papers that weren't based on hypothesis testing or collecting data; or, of course, exploration! Our suspicion is that hypotheses are often 'cooked up' in a highly proficient, and therefore quite undetectable, fashion. As we've explained, this is not difficult to do.

And the fact that hypotheses are sometimes 'cooked up' is illustrated by our final finding. Even if we err strongly on the side of caution and imagine that only half of our suspicions were justified – i.e. only half of the papers we classified as S actually contain presentational hypotheses – then our results still show that there were presentational hypotheses present in 12% of the papers that we examined. (And recall that we were careful to avoid S classifications if either of us had any doubts.) If these results held on average, then one would expect to find a presentational hypothesis or two in each issue of *The Journal of Experimental Biology* and *Journal of Biomechanics*.

[10] Clearly data collection can facilitate hypothesis testing, and be implicitly based on the desire to test hypotheses. Thus stating that the aim of some work was to collect data does not involve denying that it was to test (or help in the testing of) hypotheses.

Overall, therefore, it is reasonable to conclude that biomechanists have a bias towards presenting their research as testing hypotheses, and (especially) prefer not to present their research as if it bears no relation to hypothesis testing. Needless to say, this could be mainly pragmatic, rather than reflect widespread agreement on what counts as good scientific practice (or genuine scientific activity). If biomechanists suspect that their chances of publication (and/or funding) will be increased by presenting their work in a particular way, then many will do so even if do so is inaccurate. The practice might also have become entrenched somewhat through the apprentice-like system in which biomechanists are trained.

Selection may also, or instead, have occurred in the refereeing process. It would be interesting to discover how many of the papers we analysed were revised in response to comments from referees, and if the revisions ever led to the insertion of hypothesis testing talk that was not originally present. It would also be interesting to discover how many papers are rejected for failing to frame hypotheses for test. Alas, this would be a difficult empirical study to pull off; access to journal records, and particularly to unpublished referees' reports, would be required.

So are presentational hypotheses *really* a bad thing? Even if one rejects the view that prediction has more epistemic significance than accommodation – instead holding that how some data bears on a hypothesis is independent of when the data was collected (or when the hypothesis was considered), and even how the data was collected (provided the data is true/accurate) – it is difficult to see what advantage presentational hypotheses offer. To reject such a view is to think that whether an

author intended to test a hypothesis (and designed an experiment for that express purpose), in gathering his/her data, is scientifically irrelevant. Hence the 'window dressing' would be unnecessary unless it served to draw attention to the paper's findings in a way that could not otherwise be achieved. It is highly doubtful that it does, because one may state that some data confirms or disconfirms (or corroborates or falsifies) some hypothesis without saying anything about the intent behind the experiment to collect that data (or behind any other activity by which the data became available). So to summarise, we believe that presentational hypotheses are at best unnecessary, and at worst (often) responsible for misleading readers about the significance of research findings (or specific data).

We have still not explained precisely why hypothesis-testing seems to have become 'up-valued' in biomechanics. One reasonably plausible story, *prima facie*, is that this is due, in part, to outside philosophical influence. For example, biomechanists (perhaps as a subset of biologists) may have been looking for a way to make their work appear manifestly scientific and adopted a hypothesis-testing mode of presentation on the basis of the emphasis that this received from the likes of Popper (who is perhaps the best known philosopher of science among scientists). Alternatively, and less plausibly, perhaps many influential biomechanists became convinced that hypothesis-testing is at the heart of genuine science (and refereed the papers and proposals of others accordingly, influenced the views of their research students, and so forth). This is a fascinating historical question which awaits a definitive answer, and which would have to be the subject matter of a longer study.

Acknowledgements

Thanks to Alexander Bird, Tom McLeish, Wilson Poon, and three anonymous referees for their comments. Darrell Rowbottom's work on this paper was supported by the Templeton Foundation, as part of the 'Why ''Why?"—Methodological and Philosophical Issues at the Physics–Biology Interface' project, and subsequently by the British Academy.

References

Almbro, Maria and Cecilia Kullberg. 2008. "Impaired Escape Flight Ability in Butterflies Due to Low Flight Muscle Ratio Prior to Hibernation." *The Journal of Experimental Biology* 211: 24–28.

Anders, Christoph, Gunther Brose, Gunther O. Hofmann, and Hans-Christoph Scholle. 2008. "Evaluation of the EMG–Force Relationship of Trunk Muscles During Whole Body Tilt." *Journal of Biomechanics* 41: 333–339.

Astley, Henry C. and Bruce C. Jayne. 2007. "Effects of Perch Diameter and Incline on the Kinematics, Performance and Modes of Arboreal Locomotion of Corn Snakes (*Elaphe Guttata*)." *The Journal of Experimental Biology* 210: 3862–3872.

Barnes, Eric C. 2005. "Predictivism for Pluralists." *British Journal for the Philosophy of Science* 56: 421–450.

Bates, Mary E., Sarah A. Stamper, and James A. Simmons. 2008. "Jamming Avoidance Response of Big Brown Bats in Target Detection." *The Journal of Experimental Biology* 211: 106–113.

Chang, Chia-Lin, and Beverly D. Ulrich. 2008. "Lateral Stabilization Improves Walking in People with Myelomeningocele." *Journal of Biomechanics* 41: 1317–1323.

Clark, Andrew J. and Adam P. Summers. 2007. "Morphology and Kinematics of Feeding in Hagfish: Possible Functional Advantages of Jaws." *The Journal of Experimental Biology* 210: 3897–3909.

Donelan, J. Maxwell, David W. Shipman, Rodger Kram, and Arthur D. Kuo. 2004. "Mechanical and Metabolic Requirements for Active Lateral Stabilization in Human Walking." *Journal of Biomechanics* 37: 827–835.

Edwards, James L. 1985. "Terrestrial Locomotion without Appendages." In *Functional Vertebrate Morphology*, edited by Milton Hildebrand, Dennis M. Bramble, Karel F. Liem, and David B. Wake, 159–172. Cambridge, MA: Harvard University Press.

Ellerby, David J. and Graham N. Askew. 2007. "Modulation of Pectoralis Muscle Function in Budgerigars *Melopsitaccus Undulatus* and Zebra Finches *Taeniopygia Guttata* in Response to Changing Flight Speed." *The Journal of Experimental Biology* 210: 3789–3797.

Estrella, Sora M. and José A. Masero. 2007. "The Use of Distal Rhynchokinesis by Birds Feeding in Water." *The Journal of Experimental Biology* 210: 3757–3762.

Fleming, Alexander. 1929. "On the Bacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. Influenzae." *British Journal of Experimental Pathology* 10: 226–236.

Ford, Matthew D., Sang-Wook Lee, Stephen P. Lownie, David W. Holdsworth, and David A. Steinman. 2008. "On the Effect of Parent–Aneurysm Angle on Flow Patterns in Basilar Tip Aneurysms: Towards a Surrogate Geometric Marker of Intra-Aneurismal Hemodynamics." *Journal of Biomechanics* 41: 241–248.

Griffin, Donald R., John J. G. McCue, and Alan D. Grinnell. 1963. "The Resistance of Bats to Jamming." *Journal of Experimental Zoology* 152: 229–250.

Hanson, Norwood R. 1962. "Leverrier: The Zenith and Nadir of Newtonian Mechanics." *Isis* 53: 359–378.

Harker, David. 2008. "On the Predilections for Predictions." *British Journal for the Philosophy of Science* 59: 429–453.

Hedrick, Michael S., Robert C. Drewes, Stanley S. Hillman and Philip C. Withers. 2007. "Lung Ventilation Contributes to Vertical Lymph Movement in Anurans." *The Journal of Experimental Biology* 210: 3940–3945.

Hudson, Robert G. 2007. "What's Really at Issue with Novel Predictions?" *Synthese* 115: 1–20.

Jenkyn, Thomas R., Michael A. Hunt, Ian C. Jones, Robert Giffin, and Trevor B. Birmingham. 2008. "Toe-out Gait in Patients with Knee Osteoarthritis Partially Transforms External Knee Adduction Moment into Flexion Moment During Early Stance Phase of Gait: A Tri-Planar Kinetic Mechanism." *Journal of Biomechanics* 41: 276–283.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Marshall, Christopher D., Kit M. Kovacs, and Christian Lydersen. 2008. "Feeding Kinematics, Suction and Hydraulic Jetting Capabilities in Bearded Seals (*Erignathus Barbatus*)." *The Journal of Experimental Biology* 211: 699–708.

Melcón, Mariana L., Annette Denzinger, and Hans-Ulrich Schnitzler. 2007. "Aerial Hawking and Landing: Approach Behaviour in Natterer's Bats, *Myotis Nattereri* (Kuhl 1818)." *The Journal of Experimental Biology* 210: 4457–4464.

Parsons, Kevin J., Thilo Pfau, Marta Ferrari, and Alan M. Wilson. 2008. "High-Speed Gallop Locomotion in the Thoroughbred Racehorse. II. The Effect of Incline on Centre of Mass Movement and Mechanical Energy Fluctuation." *The Journal of Experimental Biology* 211: 945–956.

van der Merwe, Helena, B. Daya Reddy, Peter Zilla, Deon Bezuidenhout, and Thomas Franz. 2008. "A Computational Study of Knitted Nitinol Meshes for their Prospective Use as External Vein Reinforcement." *Journal of Biomechanics* 41: 1302–1309.

Medawar, Peter. 1963. "Is the Scientific Paper a Fraud?" *The Listener* 70: 377–378.

Moazen, Mehran, Neil Curtis, Susan E. Evans, Paul O'Higgins, Michael J. Fagan. 2008. "Rigid-Body Analysis of a Lizard Skull: Modelling the Skull of *Uromastyx Hardwickii*." *Journal of Biomechanics* 41: 1274–1280.

Musgrave, Alan E. 1974. "Logical versus Historical Theories of Confirmation." *British Journal for the Philosophy of Science* 25: 1–23.

Nowlan, Niamh C., Paula Murphy, and Patrick J. Prendergast. 2008. "A Dynamic Pattern of Mechanical Stimulation Promotes Ossification in Avian Embryonic Long Bones." *Journal of Biomechanics* 41: 249–258.

Popper, Karl R. 1959. *The Logic of Scientific Discovery*. New York: Basic Books.

Popper, Karl R. 1983. *Realism and the Aim of* Science. London: Routledge.

Rowbottom, Darrell P. 2008a. "Intersubjective Corroboration." *Studies in History and Philosophy of Science* 39: 124–132.

Rowbottom, Darrell P. 2008b. "The Big Test of Corroboration." *International Studies in the Philosophy of Science* 22: 293–302.

Rowbottom, Darrell P. 2011a. "Kuhn vs. Popper on Criticism and Dogmatism in Science: A Resolution at the Group Level." *Studies in History and Philosophy of Science* 42: 117–124.

Rowbottom, Darrell P. 2011b. *Popper's Critical Rationalism: A Philosophical Investigation*. London: Routledge.

Siegmund, Thomas, Matthew R. Allen, and David B. Burr. 2008. "Failure of Mineralized Collagen Fibrils: Modeling the Role of Collagen Cross-Linking." *Journal of Biomechanics* 41: 1427–1435.

Sigal, Ian A., Michael R. Hardisty, and Cari M. Whyne. 2008. "Mesh-Morphing Algorithms for Specimen-Specific Finite Element Modeling." *Journal of Biomechanics* 41: 1381–1389.

Simmons, James A., Kyler M. Eastman, Seth S. Horowitz, Michael J. O'Farrell, and David N. Lee. 2001. "Versatility of Biosonar in the Big Brown Bat, *Eptesicus Fuscus*." *Acoustic Research Letters Online* 2: 43–48.

Stewart, Michael A. (ed.) 1991. *Selected Philosophical Papers of Robert Boyle*. Indianapolis: Hackett.

Tan, Huiling, Alan M. Wilson, and John Lowe. 2008. "Measurement of Stride Parameters Using a Wearable GPS and Inertial Measurement Unit." *Journal of Biomechanics* 41: 1398–1406.

Verhulp, Eelco, Bert van Rietbergen, Ralph Müller, and Rik Huiskes. 2008. "Indirect Determination of Trabecular Bone Effective Tissue Failure Properties Using Micro-Finite Element Simulations." *Journal of Biomechanics* 41: 1479–1485.