

Northumbria Research Link

Citation: Wilkinson, Mick (2014) Clinical and practical importance versus statistical significance: limitations of conventional statistical inference. *International Journal of Therapy and Rehabilitation*, 21 (10). pp. 488-494. ISSN 1741-1645

Published by: Mark Allen Publishing

URL: <http://dx.doi.org/10.12968/ijtr.2014.21.10.488>
<<http://dx.doi.org/10.12968/ijtr.2014.21.10.488>>

This version was downloaded from Northumbria Research Link:
<http://nrl.northumbria.ac.uk/23319/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>

This document may differ from the final, published version of the research and has been made available online in accordance with publisher policies. To read and/or cite from the published version of the research, please visit the publisher's website (a subscription may be required.)

www.northumbria.ac.uk/nrl



1 Title: Clinical and practical importance versus statistical significance: limitations of conventional
2 statistical inference.

3 Submission Type: Research methodology

4 Authors: Michael Wilkinson

5 Affiliation: Faculty of Health and Life Sciences

6 Northumbria University

7 Correspondence address: Dr Michael Wilkinson

8 Department of Sport, Exercise and Rehabilitation

9 Northumbria University

10 Northumberland Building

11 Newcastle-upon-Tyne

12 NE1 8ST

13 ENGLAND

14 Email: mic.wilkinson@northumbria.ac.uk

15 Phone: 44(0)191-243-7097

16

17 Abstract word count: 193

18 Text only word count: 4727

19

20

21

22

23

24

25

26

27

28

29 Abstract

30 Decisions about support for therapies in light of data are made using statistical inference. The
31 dominant approach is null-hypothesis-significance-testing. Applied correctly it provides a procedure
32 for making dichotomous decisions about zero-effect null hypotheses with known and controlled
33 error rates. Type I and type II error rates must be specified in advance and the latter controlled by a
34 priori sample size calculation. This approach does not provide the probability of hypotheses or the
35 strength of support for hypotheses in light of data. Outcomes allow conclusions only about the
36 existence of non-zero effects, and provide no information about the likely size of true effects or their
37 practical / clinical value. Magnitude-based inference, allows scientists to estimate the 'true' / large
38 sample magnitude of effects with a specified likelihood, and how likely they are to exceed an effect
39 magnitude of practical / clinical importance. Magnitude-based inference integrates elements of
40 subjective judgement central to clinical practice into formal analysis of data. This allows enlightened
41 interpretation of data and avoids rejection of possibly highly-beneficial therapies that might be 'not
42 significant'. This approach is gaining acceptance, but progress will be hastened if the shortcomings of
43 null-hypothesis-significance testing are understood.

44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61

62 Introduction

63 The scientific method is characterised by the formulation of theories and the evaluation of specific
64 predictions derived from those theories against experimental data. Decisions about whether
65 predictions and their parent theories are supported or not by data are made using statistical
66 inference. Thus the examination of theories and the evaluation of therapies in light of data and
67 progression of 'knowledge' hinge directly upon how well the inferential procedures are used and
68 understood. The dominant approach to statistical inference is null-hypothesis-significance testing
69 (NHST). NHST has a particular underpinning logic that requires strict application if its use is to be of
70 any value at all. Even when this strict application is followed, it has been argued that the
71 underpinning 'yes or no' decision logic and the value of the 'sizeless' outcomes produced from NHST
72 are at best questionable and at worst can hinder scientific progress (Ziliak and McCloskey, 2008,
73 Batterham and Hopkins, 2006, Krantz, 1999, Sterne and Smith, 2001). The failure to understand and
74 apply methods of statistical inference correctly can lead to mistakes in the interpretation of results
75 and subsequently to bad research decisions. Misunderstandings have a practical impact on how
76 research is interpreted and what future research is conducted, so impacts not only researchers but
77 any consumer of research. This paper will clarify NHST logic, highlight limitations of this approach
78 and suggest an alternative approach to statistical inference that can provide more useful answers to
79 research questions while simultaneously being more rational and intuitive.

80

81 Scientific inference

82 With a clear picture of scientific inference, it is easier to understand the 'fit' of different statistical
83 approaches to what we wish the scientific method to achieve. Science is a way of working involving
84 formulation of theories or guesses about how the world works, calculation of the specific
85 consequences of those guesses (i.e. hypotheses about what should be observed if the theory is
86 correct), and the comparison of actual observations to those predictions (Chalmers, 1999). This
87 description of the scientific method is not contentious. How observations are used to establish the
88 'truth' of theories is more contentious. The nature of philosophy is such that there will never be
89 wholesale agreement, but a philosophy of science proposed by Sir Karl Popper is generally
90 considered an ideal to strive towards. Popper wrote that theories must make specific predictions
91 and importantly, those predictions should be potentially falsifiable through experiment (Popper,
92 1972b). It was Popper's falsifiability criteria that differentiated his philosophy from the consensus
93 approach of 'truth by verification' that predominated previously, while simultaneously overcoming
94 the problem of inductive reasoning highlighted by Scottish philosopher David Hume (Hume, 1963).
95 In short, Popper showed that it was impossible to 'prove' a theory no matter how many
96 observations verified it, but that a single contrary observation could 'disprove' or falsify a theory.
97 This thesis is often explained using the 'white swan' example. Imagine a hypothesis that all swans
98 are white. No amount of observations of white swans could prove the hypothesis true as this
99 assumes all other swans yet to be observed will also be white and uses inductive reasoning. A single
100 observation of a black (or other non-white) swan could however, by deductive reasoning, disprove
101 the hypothesis (Ladyman, 2008). In Popper's philosophy, scientists should derive *specific* hypotheses
102 from general theories and design experiments to attempt to falsify those hypotheses. If a hypothesis
103 withstands attempted falsification, it and the parent theory are not proven, but have survived to

104 face further falsification attempts. Theories that generate more falsifiable predictions and more
105 specific predictions are to be preferred to theories whose falsifiable predictions are fewer in number
106 and vague. This latter point is particularly important in relation to NHST and will be expanded upon
107 later.

108

109 Truth, variability and probability

110 Critics of Popper argue that, in reality, scientists would never reject a theory on the basis of a single
111 falsifying observation and that there is no absolute truth that more successful theories move
112 towards (Kuhn, 1996). Popper agreed and acknowledged that it would be an accumulation of
113 falsifying evidence that 'on balance of probability' would lead to the conclusion that a theory had
114 been disproven (Popper, 1972b). Herein lie two important links between statistical and scientific
115 inference, namely that *probability* must be the basis for conclusions about theories because of
116 *variability* in the results of different experiments on the same theory. Uncertainty is inescapable, but
117 statistics can allow quantification of uncertainty in the light of variability. British polymath Sir Ronald
118 Fisher first suggested a method of using probability to assess strength of evidence in relation to
119 hypotheses (Fisher, 1950, Fisher, 1973). Fisher's contributions to statistics include the introduction
120 of terms such as 'null hypothesis' (denoted as H_0), 'significance' and the concept of degrees of
121 freedom, random allocation to experimental conditions and the distinction between populations
122 and samples (Fisher, 1950, Fisher, 1973). He also developed techniques including analysis of variance
123 amongst others. He is perhaps better known for suggesting a p (probability) of 0.05 as an arbitrary
124 threshold for decisions about H_0 that has now achieved unjustified, sacrosanct status (Fisher, 1973).

125

126 Fisher's null

127 Fisher's definition of the null hypothesis was very different from what we currently understand it to
128 mean and is possibly the root cause of philosophical and practical problems with NHST that will be
129 discussed in this paper. In Fisher's work, the null was simply the hypothesis we attempt to 'nullify' or
130 in other words 'falsify' (Fisher, 1973). With this understanding, he was actually referring to what we
131 now call the 'experimental' hypothesis (denoted as H_1) and his procedures were well aligned with
132 Popper's falsification approach. The conventional zero-point null hypothesis and the procedures for
133 establishing a decision-making procedure about H_0 (i.e. retain or reject) that predominate today
134 were created by Polish mathematician Jerzy Neyman and British statistician Egon Pearson (Neyman
135 and Pearson, 1933). Despite the $p < 0.05$ being attributed to Fisher as a threshold for making a
136 decision about (his version of) H_0 , he was opposed to the idea of using threshold probabilities and
137 argued vigorously in the literature with Neyman and Pearson about this (Ziliak and McCloskey,
138 2008). Instead, Fisher argued that probability could be used as a continuous measure of strength of
139 evidence against the null hypothesis (Fisher, 1973), a point that, despite his genius, he was gravely
140 mistaken about.

141

142

143 Defining probability

144 Generally speaking, there are two interpretations of probability in statistics. The first is subjective
145 and the second objective. Subjective probability is the most intuitive and describes a personal
146 degree of belief that an event will occur. It also forms the basis of the Bayesian method of inference.
147 In contrast, the objective interpretation of probability is that probabilities are not personal but exist
148 independent of our beliefs. The NHST approach and Fisher's ideas are based on an objective
149 interpretation of probability proposed by Richard von Mises (von Mises, 1928). This interpretation is
150 best illustrated using a coin-toss example. In a fair coin, the probability of heads is 0.5 and reflects
151 the proportion of times we *expect* the coin to land on heads. However, it cannot be the proportion
152 of times it lands on heads in any *finite* number of tosses (e.g. if in 10 tosses we see 7 heads, the
153 probability of heads is not 0.7). Instead, the probability refers to an *infinite number of hypothetical*
154 coin tosses referred to as a 'collective' or in more common terms a 'population' of scores of which
155 the real data are assumed to be a sample. The population must be clearly defined. In this example, it
156 could be all hypothetical sets of 10 tosses of a fair coin using a precise method under standard
157 conditions. Clearly, 7 heads from 10 tosses is perfectly possible even with a fair coin, but the more
158 times we toss the coin, the more we would expect the proportion of heads to approach 0.5. The
159 important point is that the probability applies to the hypothetical-infinite collective and not to a
160 single toss or even a finite number of tosses. It follows that objective probabilities also do not apply
161 to hypotheses as a hypothesis in the NHST approach is simply retained or rejected in the same way
162 that a single event either happens or does not, and has no associated population to which an
163 objective probability can be assigned. Most scientists believe a p value from a significance test
164 reveals something about the probability of the hypothesis being tested (generally the null). Actually
165 a p value in NHST says *nothing* about the likelihood of H_0 or H_1 or the strength of evidence for or
166 against either one. It is the probability of data as extreme or more extreme than that collected
167 occurring in a hypothetical-infinite series of repeats of an experiment *if H_0 were true* (Oakes, 1986).
168 In other words, the truth of H_0 is assumed and is fixed, p refers to all data from a hypothetical
169 distribution probable under or consistent with H_0 . It is the conditional probability of the observed
170 data assuming the null hypothesis is true, written as $p(D|H)$.

171

172 Null-Hypothesis-Significance Testing logic

173 Based on the objective interpretation of probability, the NHST approach was designed to provide a
174 dichotomous decision-making procedure with known and controlled long-run error rates. Neyman
175 and Pearson were clear about this and in the introduction of their classic paper to the Royal Society
176 stated "... as far as a particular hypothesis is concerned, no test based on the (objective) theory of
177 probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis"
178 (Neyman and Pearson, 1933). Instead, they set about defining rules to govern decisions about
179 retaining or rejecting hypotheses such that wrong decisions would not often be made, but the
180 probability of making them in the long run would be known.

181 The starting point of the N-P approach is a pair of contrasting hypotheses (H_0 and H_1). For example,
182 H_0 could be that μ_1 (population mean ankle dorsi-flexion angle given therapy x) = μ_c (population
183 mean ankle dorsi-flexion angle given no therapy – i.e. control group), or to put it another way, the
184 difference between μ_1 and μ_c is zero. The alternative (H_1) is then generally of the form $\mu_1 \neq \mu_c$ i.e. the

185 population mean of the therapy and control groups will not be equal / will differ. Here we have the
186 first philosophical issue with the conventional use of NHST. Under the philosophy of Popper, a
187 hypothesis should be a specific prediction such that it is highly falsifiable. Popper argued a theory
188 that allows everything explains nothing (Popper, 1972a) i.e. falsifying a null of 'no difference' simply
189 allows for any magnitude of difference in any direction – hardly a severe test of a theory!
190 Furthermore, the hypothesis under consideration (i.e. a zero-effect null) is not actually the
191 hypothesis of interest, but is simply a straw man that the researcher does not believe or they would
192 not be performing the experiment. Not surprisingly, Popper was not a supporter of NHST (Dienes,
193 2008). A practical issue is also raised here. Ignoring the philosophical problem, if a null of 'no
194 difference' is rejected, a question that remains is how big is the effect? It is generally the size of
195 effect of a therapy versus a control condition / group that is of real interest, not simply that the
196 effect is different from zero in some unspecified amount and direction.

197

198 The illogic of NHST

199 Because H_0 and H_1 are mutually exclusive, if H_0 is rejected, by deduction H_1 is assumed true and vice
200 versa, if H_0 is not rejected, H_1 is assumed false. However, statistical inference and indeed science
201 does not deal in absolute proofs, truths or falsehoods, there is always uncertainty. If this uncertainty
202 is extended to this example, we have: If H_0 then *probably* NOT H_1 , data arise consistent with H_1 ,
203 therefore H_0 is *probably* false. This logic has been challenged. Pollard and Richardson (1987)
204 highlight a flaw using the following example: 'if a person is American, they are probably not a
205 member of Congress; person x is a member of Congress therefore person x is probably not
206 American'. Furthermore, Oakes (1986) points out that we are concluding the truth of H_1 based on H_0
207 being unlikely, when H_1 might be even less likely but we shall never know as it has not been tested,
208 nor has the likelihood of multiple other possible versions of H_1 . This paradox has been called the
209 fallacy of the transposed conditional (Ziliak and McCloskey, 2008)

210

211 Errors in decision making with NHST

212 NHST logic gives rise to two possible errors in decision making, namely, wrongly rejecting H_0 when it
213 is actually true (type I error) and wrongly retaining H_0 when it is actually false (type II error). To
214 correctly use NHST, the researcher chooses the acceptable risk of each error type in advance of
215 testing (subjectively and according to the type of error the researcher deems more harmful). The
216 NHST procedure ensures these error rates are fixed and controlled such that, over an infinite
217 number of hypothetical repeats of the experiment, the probability of making each type of error is
218 known (Neyman and Pearson, 1933). The probability of a type I error is termed α and is
219 conventionally and without reason set at 0.05. The probability of a type II error is termed β . This
220 error rate is less formally agreed and in the majority of research is never actually specified or
221 controlled, violating NHST decision rules. The few studies that do control β generally specify it at 0.2
222 giving the study an 80% chance ($1 - \beta$) of correctly rejecting a false H_0 , or 80% statistical power. For
223 the type II error rate to be fixed, a minimum worthwhile / interesting effect that researchers wish to
224 detect must be specified in advance of data collection, and an appropriate sample size calculated
225 that provides the specified power (and thus type II error rate). *Exactly* that number of participants

226 must be tested to control the type II error rate at the specified level. Failure to specify β in advance
227 and to control it by testing an appropriately-sized sample renders decisions about H_0 impossible
228 when it cannot be rejected i.e. was the effect really likely to be zero or was it different from zero but
229 undetectable due to low power? The only conclusion that can be drawn is one of uncertainty. Failure
230 to test the 'correct' number of participants can also result in effects not large enough to be of
231 practical / clinical importance being deemed 'significant' if a larger-than-necessary sample is used
232 (i.e. the experiment is overpowered). It should be acknowledged that, in reality, the sample sizes in
233 published studies are rarely based on an appropriate calculation. In fact, few studies would actually
234 take place if the estimated sample size had to be obtained, as they are often prohibitively large. This
235 is unlikely to change, but researchers simply need to be aware that, strictly, the decision logic of
236 NHST only applies when both error rates are actually controlled.

237

238 An example of NHST in practice

239 In the 'therapy-versus-control-group' example outlined in the previous section, having specified
240 hypotheses and error rates and calculated an appropriately-sized sample, samples (assumed to be
241 random) are taken from the hypothetical collectives of interest. The sample means for the therapy
242 (M_t) and the control (M_c) groups and the difference between them can be calculated. The standard
243 error of the mean difference (SEM_{diff}) can also be calculated. These values are used to calculate a
244 sample statistic that combines them, in this case a t statistic, where $t = (M_t - M_c) / SEM_{diff}$. In order to
245 calculate the long-run probability that such a t statistic could occur given H_0 , the hypothetical
246 collective that gave rise to this t statistic must be defined. The collective in this case is a probability
247 distribution of t statistics from an infinite number of hypothetical repeats of the experiment
248 assuming H_0 is true (so having a mean *difference* between therapy and control groups of 0 and an
249 assumed-normal distribution). This theoretical distribution represents all values of t that are
250 probable if H_0 were true. Now the decision rule is applied by defining a rejection region of the
251 distribution where t statistics are deemed so extreme that they would occur infrequently in the long
252 run if H_0 were true. The probability of obtaining a t score in that region is equal to the predefined α .
253 Thus, if the observed t from the sample data falls into the region of the probability distribution
254 beyond α , H_0 is rejected as such a t statistic would occur infrequently in the long run if H_0 were true.
255 Note that the interpretation of such a finding is that 'an effect exists in the sample that should not
256 be likely if there really was no effect in the collective sampled – therefore, there is likely to be an
257 effect larger than zero in the collective sampled. If you find this confusing, you are not alone. Little
258 can be concluded about the size of the effect of the therapy versus the control or the practical /
259 clinical value of it, which is arguably much more important (Ziliak and McCloskey, 2008, Batterham
260 and Hopkins, 2006).

261 Note that the *exact* probability of the observed t is irrelevant to the decision to reject H_0 . It need
262 only be less than α . The requirement for authors to report exact p values is actually redundant and
263 stems from a mistaken belief that p is in some way a measure of strength of evidence against H_0
264 such that the lower the p the stronger the evidence against H_0 and by extension for H_1 . It has already
265 been discussed that p is a conditional probability of the observed *data* occurring assuming a fixed H_0 ,
266 i.e. $p(D|H_0)$, as such p is not an indicator about either hypothesis. Most researchers believe the p
267 value tells them something about the probability of their hypothesis in light of the data i.e. $p(H|D)$,

268 and that the magnitude of p is in some way a continuous measure of the weight of evidence against
269 H_0 , when in fact, any given p could simply be a function of random sampling variation (Cumming,
270 2012). It is worth expanding on this point to highlight how trust in p as a form of evidence is
271 misplaced.

272 In the example provided, the t statistic for which p is calculated is derived from two random samples
273 taken from hypothetical-infinite collectives. Different samples would produce different means and
274 therefore a different t statistic and a different p value. The p value is thus a randomly-fluctuating
275 variable that can and does jump in and out of the rejection region when an experiment is repeated
276 exactly as before. Being so unreliable, how can a researcher possibly have trust in a p value as a
277 source of evidence on which to base a decision about their hypotheses? (Cumming, 2012). Note also
278 the desire for p to indicate 'magnitude' of evidence and effect. The importance of estimating the
279 likely 'size' of an effect has been recognised as the most important goal of statistical inference but a
280 p value is not it (Ziliak and McCloskey, 2008, Hopkins et al., 2009, Batterham and Hopkins, 2006).

281

282 Size matters, but so does uncertainty

283 The goal of statistical inference is to estimate likely 'true / large-sample' effects based on random
284 samples from the collective(s) of interest. However, because different samples always produce
285 different estimates, we must express the uncertainty of our estimates. This is achieved using
286 confidence intervals. The exact definition of a confidence interval is debated, but it is generally
287 accepted to be a plausible range in which the true population effect would be likely to fall with
288 repeats of the experiment. The number of repeats is infinite and 'likely to fall' refers to the
289 percentage of times a calculated interval would contain the 'true' effect (conventionally 95%). In the
290 context of the therapy-control example already used, if we repeated the experiment an infinite
291 number of times and calculated a confidence interval each time, 95% of them would contain the
292 'true' effect and 5% would not. We can never know whether the interval calculated in *this* study is
293 one of those that does or does not contain the 'true' effect. Taking a pragmatic view however, as
294 95% of all intervals *will* contain the 'true' effect then our interval is more likely to be one of those
295 that does than one of those that does not. The use of interval estimation instead of NHST is
296 becoming a necessity for publication in many journals. In fact, the International Committee of
297 Medical Journal Editors (2010) now make the following statement in their guidelines; "quantify
298 findings and present them with appropriate indicators of measurement error or uncertainty (such as
299 confidence intervals). Avoid relying solely on statistical hypothesis testing, such as P values, which
300 fail to convey important information about effect size and precision of estimates"

301

302 Suppose a 95% confidence interval for the mean difference between the control and therapy group
303 ankle dorsi-flexion angle was calculated as 1° to 10° in favour of the therapy group. A pragmatic
304 interpretation is that the intervention is likely to result in an improvement in ankle dorsi-flexion
305 range of between 1° and 10° *more* than no therapy. Assume the mean difference between the
306 sample groups was 5.5° and that the p value of a NHST (t test in this case) was < 0.05 . Using the
307 latter, the therapy would be deemed successful, with the best estimate of the 'true / large sample'
308 effect of the therapy being an improvement in ankle dorsi flexion of 5.5° . Using the confidence

309 interval, we are still confident that there is a benefit of the therapy over the control, but the size of
310 the improvement might be as little as 1° or as large as 10°. The confidence interval factors sampling
311 variability into the estimate and thus expresses the uncertainty about what the true mean difference
312 between therapy and control might be. Imagine how the discussion section of a paper might differ
313 with the NHST and confidence-interval results. The key to the discussion section in both situations
314 should be the *context* of what size of improvement in dorsi flexion is beneficial (for function, quality
315 of life, as a return for time / cost invested in the therapy etc.) (Batterham and Hopkins, 2006, Ziliak
316 and McCloskey, 2008). Given these factors, the conclusion might be that the benefit of the therapy is
317 uncertain as it could be hardly beneficial at all or extremely beneficial. If a worthwhile improvement
318 was say 1°, then the therapy is more than likely to be worthwhile, but if an improvement is not
319 worthwhile or beneficial unless it exceeds 5°, then the conclusion will be less favourable. Note that
320 for a therapy to be deemed 'successful' with *reasonable likelihood* using this approach, two
321 conditions must be satisfied: 1) the confidence interval must exclude zero and; 2) the lower limit of
322 the confidence interval must be equal to or greater a smallest clinically worthwhile / beneficial
323 effect. Both conditions can lead to very conservative conclusions (Atkinson and Nevill, 2001).
324 Suppose that in the above example, a smallest-beneficial improvement was deemed to be 3°, the
325 conclusion would have to be that the therapy is possibly of no use even though it could be very
326 useful indeed. If the interval ran from -1° to 10°, again the conclusion would be unfavourable
327 because not only might the true effect be no effect at all, it might also result in a worsening of ankle
328 range, though the true effect could still be a gain in dorsi-flexion of up to a 10°, and more of the
329 interval lies above the smallest-beneficial effect than below it! There is a real danger of throwing the
330 baby out with the bath water. Clearly, what is required is a method that allows calculation of the
331 likelihood of the true effect in relation to a smallest effect of clinical / practical importance.
332 Researchers and in particular clinical practitioners make these judgements subjectively anyway, so
333 why not incorporate them into statistical inference. There is a method for doing just this (Batterham
334 and Hopkins, 2006, Hopkins et al., 2009).

335

336 Magnitude-based inference

337 Magnitude-based inference (MBI) involves calculating the chances (probability) that the 'true' effect
338 exceeds or is less than an *a priori* determined smallest-worthwhile / clinical or practically-important
339 effect. Proponents of this approach argue that the criteria of reasonable certainty adopted in the
340 confidence interval approach is too stringent and can result in conclusions of therapies /
341 interventions being non beneficial when in fact the likelihood of them being practically / clinically
342 worthwhile are extremely high (Batterham and Hopkins, 2006). Many researchers feel
343 uncomfortable making subjective decisions about the value of the smallest worthwhile effect and
344 argue that hypothesis testing is more scientific and objective. However, it was pointed out earlier
345 that estimating the sample size to control type II error in a NHST-based study, requires an estimate
346 of the smallest-worthwhile effect that is no less subjective. The choice of the smallest beneficial /
347 clinically worthwhile effect size provides the crucial context against which the results of the study
348 should be interpreted.

349 Making inferences about 'true' effect magnitudes based on smallest-worthwhile effects facilitates
350 more enlightened interpretations of data, though the process and the interpretation of the outcome

351 can be more challenging. It is often easier to “inspect the p value.....and then declare that either
352 there is or there is not an effect” (Batterham and Hopkins, 2006, p.56), but to do so might prevent
353 new knowledge about practically beneficial therapies and restrain progress of research. Lack of
354 knowledge of MBI need not be a barrier as the tools and instructions required to perform, interpret
355 and present such analyses are readily available and simple to apply (Hopkins, 2000). The MBI
356 approach is also congruent with Popper’s falsification approach whereby the expected effect of a
357 therapy is also the smallest worthwhile / important effect for the therapy to be supported. Note that
358 the smallest-worthwhile effect is precise and is the effect of interest (rather than a zero-effect null),
359 consistent with the falsification approach (Popper, 1972a).

360 The MBI approach is gaining acceptance in some of the most popular periodicals for sport, exercise
361 and medicine-related research (Hopkins et al., 2009, May et al., 2007, Hopkins et al., 2011). For a
362 therapy-related example of the application of MBI in a published study, readers are referred to May
363 *et al.* (2007) which examined the effectiveness of anti-inflammatory gel application in the treatment
364 of wrist tenosynovitis in kayakers. In addition to the fictional ankle dorsi flexion example described
365 above, the study helps to further demonstrate the practical application of the MBI approach.

366

367 Conclusions

368 Significance testing is a procedure for making black and white decisions about zero-effect null
369 hypotheses with known and controlled long-run error rates. Type I and type II error rates must be
370 specified in advance and a required sample size calculated and tested to ensure type II error rate is
371 controlled at the specified level. The outcome allows conclusions about the likely existence of non-
372 zero effects but provides no information about the likely size of true effects or their practical /
373 clinical value. The approach is also at odds with accepted philosophies of science.

374 To estimate the true size of an effect and its likelihood in relation an effect magnitude of practical /
375 clinical importance, magnitude-based inference provides the solution. The approach is gaining
376 acceptance and progress will be hastened if researchers appreciate the shortcomings of traditional
377 NHST. It is recommended that researchers begin to incorporate the subjective-clinical judgements
378 commonly made in light of experimental data, and expressions of uncertainty, into their inferential
379 statistical analysis. This will ensure more considered and enlightened interpretations of data and
380 avoid discounting possibly highly practically / clinically beneficial treatments because they are not
381 statistically significant.

382

383 Key points

384 Even used properly, NHST only gives yes / no decisions about zero-effect null hypotheses which are
385 always false and of no interest anyway.

386 Estimates of the size of true / large sample effects in NHST do not encompass uncertainty due to
387 sampling variation.

388 Outcomes of NHST are without context of what is clinically / practically important.

389 Statistical inference should estimate the likely size of true effects using confidence intervals.
390 Confidence intervals should be interpreted relative to *a priori* specified clinically / practically
391 worthwhile effect magnitudes.
392 Probabilities of true effects in relation to clinically / practically worthwhile effects should form the
393 basis for interpretation of experimental data.

394

395 References

- 396 Atkinson G & Nevill AM (2001). Selected issues in the design and analysis of sport performance
397 research. *J Sports Sci*, **19**, 811-827.
- 398 Batterham AM & Hopkins WG (2006). Making meaningful inferences about magnitudes. *Int J Sports*
399 *Physiol Perf*, **1**, 50-57.
- 400 Chalmers AF (1999). *What is this thing called science?*, Buckingham, Open University Press.
- 401 Cumming G (2012). *Understanding the new statistics: Effect sizes, confidence intervals and meta*
402 *analysis*, New York, Taylor and Francis Group.
- 403 Dienes Z (2008). *Understanding Psychology as a Science: an introduction to scientific and statistical*
404 *inference*, Basingstoke, Palgrave Macmillan.
- 405 Fisher R (1950). *Statistical methods for research workers.*, London, Oliver and Boyd.
- 406 Fisher R (1973). *Statistical methods and scientific inference*, London, Collins Macmillan.
- 407 Hopkins WG. 2000. *A new view of statistics. Internet Society for Sport Science* [Online]. Available:
408 <http://www.sportsci.org/resource/stats/>.
- 409 Hopkins WG, Batterham AM, Impellizzeri FM, Pyne DB & Rowlands DS (2011). Statistical perspectives:
410 all together NOT. *J Physiol*, **589**, 5327-5329.
- 411 Hopkins WG, Marshall SW, Batterham AM & Hanin J (2009). Progressive statistics for studies in
412 sports medicine and exercise science. *Med Sci Sports Exerc*, **41**, 3-12.
- 413 Hume D (1963). *An enquiry concerning human understanding*, Oxford, Oxford University Press.
- 414 Icmje. 2010. *Uniform requirements for manuscripts submitted to biomedical journals* [Online].
415 Available: [http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-](http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html#)
416 [submission.html#](http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html#) [Accessed 16/09/14 2014].
- 417 Krantz DH (1999). The null hypothesis testing controversy in psychology. *J Am Stat Assoc*, **94**, 1372-
418 1381.
- 419 Kuhn TS (1996). *The Structure of Scientific Revolutions*, Chicago, The University of Chicago Press.
- 420 Ladyman J (2008). *Understanding philosophy of science*, Oxford, Routledge.
- 421 May JJ, Lovell G & Hopkins WG (2007). Effectiveness of 1% diclofenac gel in the treatment of wrist
422 extensor tenosynovitis in long distance kayakers. *J Sci Med Sport*, **10**, 59-65.
- 423 Neyman J & Pearson ES (1933). On the problem of the most efficient tests of statistical hypotheses.
424 *Philos Trans R Soc Lond A*, **231**, 289-337.
- 425 Oakes M (1986). *Statistical inference: A commentary for the social and behavioural sciences*, New
426 Jersey, Wiley.

- 427 Pollard P & Richardson JTE (1987). On the probability of making type I errors. *Psychol Bull*, **102**, 159-
428 163.
- 429 Popper KR (1972a). *The Logic of Scientific Discovery*, London, Hutchinson & Co Ltd.
- 430 Popper KR (1972b). *Conjectures and Refutations: The Growth of Scientific Knowledge*, London,
431 Routledge and Kegan Paul Ltd.
- 432 Sterne JaC & Smith GD (2001). Sifting the evidence - what's wrong with significance tests? *BMJ*, **322**,
433 226-231.
- 434 Von Mises R (1928). *Probability, Statistics and Truth*, London, Allen and Unwin.
- 435 Ziliak ST & McCloskey DN (2008). *The Cult of Statistical Significance: how the standard error costs us*
436 *jobs, justice, and lives*, Michigan, University of Michigan Press.
- 437