Magnitude-based inference and its application in user research

Abstract. Magnitude-based inference offers a theoretically justified and practically useful approach in any behavioural research that involves statistical inference. This approach supports two important types of inference: mechanistic inference and practical inference to support real-world decision-making. Therefore, this approach is especially suitable for user research. We present basic elements of magnitude-based inference and examples of its application in user research as well as its merits. Finally, we discuss other approaches to statistical inference and limitations of magnitude-based inference, and give recommendations on how to use this type of inference in user research.

Keywords: user research; quantification; statistics; inference; usability testing; user-experience

Highlights

- Magnitude-based inference is a useful alternative for analysing user-research data.
- Goal-setting in user research is supported by choosing a smallest important effect.
- The approach uses the smallest important effect in making an inference.
- As a consequence, a clear effect is never an artefact of sample size.
- Practical inference is supported by weighing harm and benefit appropriately.

Contents

“It’s better to observe than to criticise.” (R.C. Wellins, personal communication, 13/2/2011)

“Best of all is to convey the magnitude of the effect and the degree of certainty explicitly.” (Pinker, 2014, p. 45)

“Usually what one wants to know is not whether the change makes *any* difference, but to know how likely it is that the change will be big enough.” (Landauer, 1997, p. 222)

1. Introduction

A researcher conducts a study comparing two software designs in terms of their usability. She conducts usability tests with two groups, each using one of the designs, and collects various measures. These include perceived usability, error rate and time-on-task. The researcher then compares the two groups in terms of their mean scores on the measures, using a *t* test. She finds that, although differences in mean scores are apparent, the test results do not show statistical significance. What should the researcher conclude about the difference in usability between the two designs?

Statistical inference is common in user research, and more generally in human-computer interaction and the behavioural sciences. The null hypothesis is a statement of the absence of the effect that is being tested, for example the difference in mean scores between two groups is 0. Typically, this hypothesis is tested to statistically demonstrate an effect. Sometimes, confidence intervals are added to provide more information or as an equivalent to (or surrogate for) the test results. The *aim* of this paper is to be translational by theoretically making the case for an

alternative approach, called magnitude-based inference, with several benefits for research in human-computer interaction, and by empirically illustrating this approach and its advantages, with examples from user research. This approach has been implemented and used extensively in the sport and exercise sciences and is therefore not new. However, we demonstrate that the approach is equally applicable in other domains such as user research, and human-computer interaction and behavioural research more generally; therefore, the use of the approach outside of sport and exercise is new. To facilitate understanding, we contrast this approach with the traditional approach of testing the null hypothesis and use illustrative examples from user research. Perhaps surprisingly, we are not advocating that researchers abandon the existing practice of analysing their data through tests of the null hypothesis with common statistical packages, but rather that they augment their existing practice by making more informative use of the results through magnitude-based inference. In particular, the results that researchers already routinely produce can be used as input for magnitude-based inference in ready-made spreadsheets that are publicly available on the Internet. To reiterate, we do not claim to present a completely new method or approach, but make the case for and demonstrate the benefits of using a recently developed approach in sport and exercise science to a new domain: user research (and human-computer interaction more widely). In this sense, this work aspires to be translational.

After a brief introduction of quantification in user research in the next section, we discuss the existing practice of testing the null hypothesis in Section 3 and present magnitude-based inference as an attractive alternative in Section 4. Section 5 provides illustrations of the application of magnitude-based inference to further demonstrate its advantages. After discussing other approaches to inference

(Section 6), we discuss limitations of magnitude-based inference (Section 7) and present recommendations for its use (Section 8).

2. Quantification in user research

The term ‘user research’ encompasses various activities such as usability testing and user-experience testing (Sauro & Lewis, 2012). This work studies the quality of the interaction between human users and interactive artefacts (computers, but also other devices, systems and services) in leisure and at work. More specifically, user research has been defined as “the systematic study of the goals, needs, and capabilities of users so as to specify design, construction, or improvement of tools to benefit how users work and live” (Schumacher, 2009, p. 6).

Following previous work in education (Scriven, 1967) and focusing on usability research, Grossman et al. (2009) distinguish between formative and summative research. In usability research, usability is measured as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, in a specified context of use” (ISO, 1998, p. 2). Typical measurements include psychometric data (e.g. usability- or user-experience questionnaire data), error rate and time-on-task. In formative usability research, users’ interaction with an artefact is studied to generate data that, when analysed, provide information to inform system improvement. Summative research establishes the quality interaction of an artefact in comparison with another artefact or a benchmark. Sauro and Lewis’s (2012) first few chapters and this paper focus on quantitative inference in summative research.

In particular for summative research, the use of the ‘gold-standard’ design for causal inference, the so-called experimental design or experiment (Cairns & Cox, 2008;

Purchase, 2012; Hornbæk, 2013), is recommended where appropriate (Lazar et al., 2010). This is because this type of design allows researchers to manipulate one or more the factors (independent variables, e.g., the usability of a website design) and observe the effects on quantitative measures, with the units of observation (human research participants) randomly assigned to treatments (e.g., website designs that differ in usability). Although quasi-experimental designs involve the manipulation of one or more independent variables, these designs lack random assignment of units to treatments. Because of this lack of control, causal inference is more difficult and, some will argue, impossible (Lazar et al., 2010). Correlational (or non-experimental; Lazar et al., 2010) designs have neither manipulation nor random assignment and are therefore the weakest designs in terms of causal inference. User researchers normally employ techniques from inferential statistics to draw conclusions from the data that they have collected, based on null-hypothesis significance-testing (NHST).

3. Statistical inference in user research

Sauro and Lewis (2012) and other human-computer interaction researchers (Landauer, 1997; Lazar et al., 2010) provide recommendations for statistical inference in user research. The null hypothesis is tested statistically. If the probability ('*p*-value') of the test result under the null hypothesis is smaller than the significance level (usually set at 0.05 or 5%) then the researcher rejects the null hypothesis and thereby concludes that there is an effect (e.g., the design of the websites that were compared in the research has an effect on users' time-on-task). NHST can be and has been applied to experimental, quasi-experimental and correlational designs, although the dominant view is that only the results of experimental designs allow causal inference.

NHST is supplemented with confidence intervals and sample size estimation for NHST. Confidence intervals are used to show the range of plausible values of the test statistic in the population (e.g., the likely range of the difference in mean score between two groups) and to infer whether there is a statistically significant effect. For example, with a mean difference of 10 points in usability scores (using the System Usability Scale [SUS]; Sauro, 2011), the 95%-confidence interval of the mean difference may have a lower limit of 5 and an upper limit of 15. As this interval does not include 0, the difference in means is statistically significant at the 5%-level. In this inference, confidence intervals are used as an equivalent (or surrogate) technique for testing the null hypothesis.

According to recommendations in the human-computer interaction literature (e.g., Landauer, 1997; Wilkinson et al., 1999; Cairns & Cox, 2008; Kaptein & Robertson, 2012; Purchase, 2012; Hornbæk, 2013) and elsewhere (Wilkinson et al., 1999), effect sizes and descriptives should be reported as part of the results of NHST. However, actually achieved effect sizes are rarely reported (Hornbæk et al., 2014).

Prospective power analysis is conducted to estimate the required sample size. This is for a researcher to have a sufficient chance (e.g., 0.80 or 80%) to detect an effect of a particular size, if it exists, in the population from which a sample has been drawn in the study. Lenth (2006-9) recommends that power analysis should be done prospectively rather than retrospectively and the analysis should be based on practically important effect sizes. Again, this technique of sample size estimation is based on NHST.

4. Magnitude-based inference

In this section we theoretically make the case for magnitude-based inference as an alternative to NHST by introducing the concepts of mechanistic and practical significance within magnitude-based inference as well as sample size estimation for both of these and by presenting its merits. The following quotation from human-computer interaction can be used as one of several motivations for considering the use of magnitude-based inference over NHST: “usually what one wants to know is not whether the change makes *any* [emphasis in original] difference, but to know how likely it is that the change will be big enough” (Landauer, 1997, p. 222; see also Drury, 2015).

4.1. Inference of mechanistic and practical significance

4.1.1. Mechanistic inference

Hopkins (2007) distinguishes two types of inference as alternatives to statistical significance (according to NHST): mechanistic inference and practical (‘clinical’) inference. Both use the probabilities of three ranges of the obtained effect as a basis for inference, but the two differ in their inference rules. Mechanistic inference is used to test an effect irrespective of its practical application, to which we turn now.

For descriptive purposes, an effect can be classified in terms of its size as positive, trivial or negative. A positive effect falls above the threshold of the smallest important positive effect that is defined by the researcher. A negative effect falls below the threshold of the smallest important negative effect that is defined by the researcher. The size of a trivial effect lies between the thresholds of the smallest important negative and positive effects.

For inference proper, the chances of an effect being positive, negative or trivial are used. The chances of the effect being positive are defined as those of the effect falling above the threshold of the smallest important positive effect. The chances of the effect being negative are defined as those of the effect falling below the threshold of the smallest important positive effect. The chances of a trivial effect are defined as 100% minus the sum of the chances of a positive effect and the chances of a negative effect (effectively these are the chances of the effect lying between the two thresholds).

An inference is then made from the chances of each of three ranges of outcome (positivity, triviality and negativity) as follows. An unclear effect is one that occurs if both the chances of the obtained effect being positive (in relation to the smallest important positive effect) and the chances of the effect being negative (in relation to the smallest important negative effect) are too large (e.g., both greater than the default value of 0.05 or another cut-offs that is deemed appropriate for positivity and negativity). Otherwise, the effect is clear, seen as substantially positive, negative or trivial and considered to have the size of the observed value, with a qualification of probability (see Table 1).

As an illustration, Figure 1 (top, mechanistic inference) shows chances for the three ranges of the obtained effect size (positivity, triviality and negativity) “that the true value will have the observed magnitude of the outcome statistic” (Batterham & Hopkins, 2006, p. 54), cut-offs of chances and inference through qualitative descriptors for mechanistic inference (from Table 1). For example, the first result in Figure 1 under ‘mechanistic inference’ shows chances of 0.01% of the effect being negative, 0.99% of the effect being trivial and 99% of the effect being positive. The qualitative descriptors from Table 1 are then applied to infer that the effect is very

likely positive (between 95% and 99.5%). For another example, the third result shows chances of 2% of the effect being negative, 33% of the effect being trivial and 65% of the effect being positive. The qualitative descriptors from Table 1 are then applied to infer that the effect is possibly positive (between 25% and 75%). For a final example, the last result shows chances of 7% of the effect being negative, 36% of the effect being trivial and 57% of the effect being positive. Because the chances of the effect being positive and the chances of the effect being negative both exceed 5% the inference is that the result is unclear; the advice would be to collect more data in order to achieve a clear result.

_____ Insert Table 1 about here. _____

_____ Insert Figure 1 about here. _____

4.1.2. Practical inference

Practical inference is used to test an effect that has a practical application. For descriptive purposes, an effect can be classified in terms of its size as beneficial, negligible or harmful. A beneficial effect falls above the threshold of the smallest important beneficial effect that is defined by the researcher.[1] A harmful effect falls below the threshold of the smallest important harmful effect that is defined by the researcher. The size of a negligible effect lies between the thresholds of the smallest important harmful and beneficial effects.

[1] This is true if a beneficial effect is defined as positive, for example an increased hit rate on the improved version of a website compared to the existing version. If the a beneficial effect is defined as negative, for example decreased time-on-task on the improved version of a website compared to the existing version, then a beneficial effect would be defined as falling below the (negative) threshold of the smallest important beneficial effect. Along similar lines, a similar caveat applies to the definition of a harmful effect.

For inference proper, the chances of an effect being beneficial, harmful or negligible are used. The chances of the effect being beneficial are defined as those of the effect falling above the threshold of the smallest important beneficial effect. The chances of the effect being harmful are defined as those of the effect falling below the threshold of the smallest important harmful effect. The chances of a negligible effect are defined as 100% minus the sum of the chances of a beneficial effect and the chances of a harmful effect (effectively these are the chances of the effect lying between the two thresholds).

Inference of the practical importance of an effect is based on the notions of Type-1 practical ('clinical') error and Type-2 practical error. The concept of Type-1 practical error is analogous to that of Type-I error in NHST (rejecting the null hypothesis when it is true); the concept of Type-2 practical error is analogous to that of Type-II error in NHST (retaining the null hypothesis when it is false). In practical inference, Hopkins' (2007) underlying idea is that, in the practical ('clinical') application of effects, the chance of using a harmful effect (a Type-1 practical error) needs to be far smaller than that of not using a beneficial effect (a Type-2 practical error), irrespective of the exact size of these chances.

An inference is then made from the chances of each of three ranges of outcome (benefit, negligibility and harm) as follows. If the chances of the effect being beneficial (in relation to the smallest important beneficial effect) are greater than the suggested cut-off of 25% for a Type-2 practical error and the chances of the effect being harmful (in relation to the smallest important harmful effect) are greater than the suggested cut-off of 0.5% for a Type-1 practical error then the effect is unclear. If the chances of benefit are greater than 25% and the chances of harm are smaller

than 0.5% then the effect is clearly beneficial. Otherwise, the effect is clearly negligible or harmful.

As an illustration, Figure 1 (bottom, practical inference) shows chances for the three ranges of the obtained effect size (benefit, negligibility and harm) that the true value will have the observed magnitude of the outcome statistic, cut-offs of chances and inference through qualitative descriptors for practical inference (from Table 1). For example, the first result in Figure 1 under 'mechanistic inference' shows chances of 0.01% of the effect being harmful, .99% of the effect being negligible and 99% of the effect being beneficial. The qualitative descriptors from Table 1 are then applied to infer that the effect is very likely beneficial (between 95% and 99.5%) and the recommendation is to use the effect (for example, to choose the new improved version of a website over the existing version). For another example, the third result shows chances of 2% of the effect being harmful, 33% of the effect being negligible and 65% of the effect being beneficial. Because the chances of harm exceed the cut-off for a Type-1 practical error (.5%) even though the chances of benefit exceed the cut-off for a Type-2 practical error (25%), the effect is deemed unclear and the recommendation is not to use the effect (for example, not to choose the new improved version of a website over the existing version). Note that here mechanistic inference (clear and 'possibly positive') and practical inference ('unclear; do not use') differ. This is because the two types of inference use different inference rules. For a final example, the last result shows chances of 7% of the effect being harmful, 36% of the effect being negligible and 57% of the effect being beneficial. Again, because the chances of harm exceed the cut-off for a Type-1 practical error (.5%) even though the chances of benefit exceed the cut-off for a Type-2 practical error (25%), the effect is deemed unclear and the recommendation is not to use the effect.

In user research and human-computer interaction more generally, as in other domains, practical inference would be attempted when benefit and harm can be defined. In the context of sport and exercise science, an intervention would be deemed harmful were it to cause a decrease in typical sporting performance/fitness. Harm may also be evident when comparing two different interventions. For example, a new training technique designed to improve speed may result in a smaller improvement in performance than the traditional approach to speed training – in this regard the new training would be considered harmful.

Examples of outcomes of human-computer interaction with benefit include faster and more accurate task performance as a result of better designed human-computer interfaces (Nielsen, 1993). Instances of outcomes with harm include infection by malware as a result of breached computer security when users act erroneously in response to malware warnings (Atzeni et al., 2014), patients' death as a result of healthcare staff erroneously entering numbers on medical devices (Oladimeji et al., 2011; Wiseman et al., 2013), violations of privacy by Internet-enabled robots for personal or domestic use when these robots share "the information required for object recognition, navigation and task completion in the real world" (Pagallo, 2013, p. 501) and financial, physical and psychological privacy threats as a result of people using wearable sensors that continuously capture physiological data and use these to infer "the wearer's behavior and psychological state in realtime" (Raij et al., 2011, p. 12). As in other fields (e.g. sport and exercise science), in human-computer interaction, a negligible effect can be defined as one that is neither beneficial nor harmful.

Hopkins (2007) suggests using cut-off points for % chances of 0.5% for harm and 25% for benefit.[2] As an alternative decision rule, Hopkins et al. (2009) suggest researchers should conclude that an effect is clear if the odds of benefit are 66 times greater than the odds of harm or vice versa; otherwise, the result is unclear. The justification for the choice of an odds ratio of 1:66 as a cut-off is that this corresponds with the suggested default probabilities of 0.005 for harm and 0.25 for benefit. As an illustration, Figure 1 (bottom) shows chances for the three ranges of effect size (harm, negligibility and benefit) that the true value will have the observed magnitude of the outcome statistic, probability cut-offs and inference through qualitative descriptors for practical inference. This further demonstrates the richness of magnitude-based inferences.

For reporting results in a publication, Hopkins (2007) recommends presenting 90%-confidence intervals and mechanistic inference for all effects. Batterham and Hopkins (2006) recommend using a 90% level of confidence to avoid the CI being used as a surrogate NHST. This choice is deemed appropriate, as the authors consider the chances of the true value higher than the upper limit or lower than the lower limit both at 0.05 as 'very unlikely' (see Table 1). For effects that have a direct practical application, practical inference should also be reported. For instance, when a new interaction technique is compared with an existing interaction technique it may be possible to define harm and benefit in terms of speed and accuracy, with the existing method as a baseline, and practical inference would be appropriate. However, if two methods are compared and neither can be seen as a baseline, harm and benefit would not be defined, so practical inference would not be attempted.

[2] Although the suggested defaults are well argued by Hopkins (2007), they are not 'set in stone' and researchers can define their own probabilities for harm and benefit, based on their own judgement.

Mechanistic and practical inference will often, but not always, be consistent with each other. The reason why any differences are unavoidable is that it is recommended in mechanistic inference for thresholds for substantially positive and negative effects to be equal, but in practical inference they should be unequal, as the chance of using a harmful effect should be much smaller than the chance of not using a beneficial effect.

4.1.3. Smallest important effect

The approach of magnitude-based inference forces a researcher to specify the minimum mechanistically and practically important effect, whereas in NHST the data are normally tested against an unrealistic null effect. The researcher can select an effect size based on theoretical or practical considerations. Within the sport and exercise science literature, the smallest worthwhile effect can be determined in one of three ways. Firstly, from years of practical experience of working with a sport, a researcher can develop a sound belief of what change in performance is needed to justify putting their athletes through an intervention. Secondly, the recent quantification and publication of the within-athlete variability between performances in several different sports provides researchers with important information of the smallest effect needed to detect a worthwhile change in performance. For example, the between-competition variability for top junior swimmers is ~1%; therefore, any strategy to improve performance needs to be at least 0.5 of this variability (Stewart & Hopkins, 2000). Finally, in the absence of a priori belief or statistical quantification of what would constitute the smallest worthwhile change in performance, inferences can be based on proposed standardised thresholds for small, moderate and large changes of 0.2, 0.6 and 1.2 *SD*s (Hopkins et al., 2009). Still, researchers can define their own thresholds for effect size, based on their own judgement.

In user research and human-computer interaction more generally, the choice of a smallest mechanistically or practically relevant effect may be based on different considerations. First, a history of practical experience working in a particular area, such as website usability, may indicate what change in, for example, a usability metric, is needed to justify a design- or purchase decision (Landauer, 1997), for example a 10% improvement in touchscreen accuracy. Second, consistent with principles of usability engineering (Wixon, 2011), a research team may set a measurable product goal in a process of continual improvement as the smallest required change that is considered important (e.g., a 20-point difference on the SUS – Sauro & Lewis, 2012, p. 71; Sauro, 2011, p. 125; or a or 1-standard deviation difference – Sauro, 2011, p. 120).

4.1.4. Magnitude-based inference for designs and contexts in human-computer interaction

Studies with three types of design – experimental, quasi-experimental and correlational – can benefit from the application of magnitude-based inference, each time by using the results from NHST as input for inference. For instance, the following results of an unrelated *t* test for comparing two website designs in terms of perceived usability, as measured by the SUS, can be entered into the appropriate spreadsheet for raw difference between means and other *t*-distributed effect statistics: *p*-value, mean difference, degrees of freedom, confidence level and smallest important effect (e.g., a 20-point difference on the SUS). The main results from the spreadsheet include chances for the three ranges of effect size (negative/harmful, trivial/negligible and positive/beneficial) and odds of benefit to harm.

Magnitude-based inference can also be applied in various contexts within human-computer interaction. In the following examples, bear in mind Hopkins' (2007) advice to use mechanistic inference for all effects and, additionally, practical inference for effects that have a direct practical application. Lab-based experiments (except usability tests) will normally use mechanistic inference only. As an example, the effect of user-interfaces on task performance in a lab environment (Komarov et al., 2013) would be tested using *t* tests on specific comparisons of interest between interfaces, with the results as input for magnitude-based inference on the difference between means in terms of the chances that the effect tested in each comparison was positive, trivial or negative (e.g. "likely positive"). As in Komarov et al.'s (2013) study, field experiments and usability tests will normally use both mechanistic and practical inference. Therefore, in addition to mechanistic inference, the results of *t* tests would also be used as input for magnitude-based inference on the difference between means in terms of the chances that each effect was beneficial, negligible or harmful and, based on this, a recommendation for practical use for each effect would follow (e.g., "very likely beneficial – use"). Correlational studies will normally use mechanistic inference only. As an example, the relation between personality and user-experience (De Oliveira et al., 2013) can be tested using Pearson's correlation, with the results as input for magnitude-based inferences on the correlation in terms of the chances that each correlation was positive, trivial or negative (e.g., the correlation coefficient shows a likely small, possibly moderate etc. relationship).

4.1.5. Software support

Various spreadsheets (*http://www.sportsci.org/*) have been developed to support mechanistic and practical inference for numerous statistics. In one of these spreadsheets (*xcl.xls*), the inference is based on the following input: a *p*-value, the

value of the effect size statistic (e.g., mean difference), degrees of freedom (where applicable), confidence level, and threshold values for a beneficial or positive effect and for a harmful or negative effect. The required input to the spreadsheet can be obtained as output from statistical analysis conducted with a common statistics package, or a general-purpose package with statistical functions, such as a spreadsheet program.

4.1.6. Illustrative example from sport- and exercise science

Here, we provide an example of the application of magnitude-based inferences within the domain of sport and exercise science. A sports researcher is interested in whether a new, commercially available nutritional supplement has a beneficial or harmful effect on elite cyclists' 40 km time-trial performance – the faster the time, the better the performance. The researcher conducts an experiment to examine the effect of two different doses of the supplement, a low dose and a high dose. An experimental crossover design is used whereby all of the cyclists perform the time trial under three different conditions (placebo [no supplement], low dose and high dose), in a counterbalanced manner. The researcher's experience led to the belief that the smallest worthwhile change in 40 km time-trial performance was -1%. The mean (± *SD*) performance times were 59.5 ± 1.6 min (low dose), 60.9 ± 2.2 min (high dose) and 60.5 ± 1.9 min (placebo), and the application of magnitude-based inferences enabled the sports researcher to calculate the % chances of benefit (or harm), with reference to a change of -1%. The following inferences are based on the results of *t* tests, conducted with a common statistics package, which are subsequently used as input into a spreadsheet (*xcl.xls*) for magnitude-based inference (available at *http://www.sportsci.org/*). In this example, compared to placebo, the low dose improved performance by -1.7% (90%-confidence interval -2.4

to -0.9%), with a 92% chance of benefit and 0.0% chance of harm; a low dose of the supplement is therefore likely to be beneficial and recommended. However, compared to placebo, the high dose impaired performance by 0.7% (90%-confidence interval -0.1 to 1.5%), with a 0% chance of benefit and a 25% chance of harm; a high dose of the supplement is therefore most unlikely beneficial and not recommended.

4.2. Sample size estimation

Corresponding with Hopkins' (Batterham & Hopkins, 2006; Hopkins, 2007) two types of magnitude-based inference from sample data, Hopkins (2006a) has developed two techniques for estimating the required sample size for empirical studies. These techniques are used instead of the common technique of power analysis to estimating sample size based on NHST. Typically, the required sample sizes in the new techniques are two to three times smaller than those in the old technique.

The new techniques "are based on (a) acceptable error rates for a […] practical decision arising from the study and (b) adequate precision for the effect magnitude" (Hopkins, 2006a, p. 65). The first technique (for [a]) estimates the required sample size for practical inference. The minimum sample size is selected such that, given the smallest harmful and beneficial important effects specified by the researcher, a particular 'decision' value of the effect size is identified. For this effect size the probability of a Type-1 practical error (using an effect as beneficial when real-world application is harmful) is 0.005 and the probability of a Type-2 practical error (choosing not to use an effect that in real-world application is beneficial) is 0.25. The second technique (for [b]) estimates the required sample size for mechanistic inference. The minimum sample size is selected such that, given the smallest mechanistically important effect specified by the researcher, the confidence interval does not overlap substantial positive and negative values.

4.3. Benefits of magnitude-based inference

Magnitude-based inference has been developed in sport and exercise science (e.g., Batterham & Hopkins, 2006; Hopkins, 2007; Hopkins et al., 2009), supported by spreadsheets freely available on the Internet. This approach is increasingly being applied, as shown by a rapid growth in publications using this approach and by citations (for Hopkins et al. [2009] from 8 citations in 2009 to 214 in 2014, with more than 740 citations in total). Research using this approach has been published in many of the leading journals in sport and exercise science (e.g., Medicine and Science in Sports and Exercise, Sports Medicine, Journal of Science and Medicine in Sport, and International Journal of Sports Physiology and Performance).

Magnitude-based inference is particularly suited for user research, as it can provide clearly interpretable and relevant results. However, until now this type of inference has remained unfamiliar in user research and human-computer interaction. Based on our presentation of magnitude-based inference in the previous sections, we now present some of its main features and benefits.

1 Requires the researcher to define a smallest important effect. Instead of testing a point hypothesis (Murphy & Myors, 1999), magnitude-based inference makes inferences with respect to three ranges of the effect size in relation to the smallest important effect (see Figure 1): negative, trivial and positive in mechanistic inference or harmful, negligible and beneficial in practical inference (Batterham & Hopkins, 2006). The purpose of the inference of mechanistic importance is to decide on the existence of a cause-effect or correlational relationship between variables, irrespective of its practical importance. The purpose of the inference of practical importance is to aid real-world decision-making regarding the use or implementation

of, for example, a new training method (in sport and exercise) or a new product design (in user research), irrespective of mechanistic importance.

2 Uses the smallest important effect size, together with the observed effect, as an integral part of inference. As a consequence, inferences are not an artefact of sample size. For example, in mechanistic inference a trivially small effect (smaller than the smallest important effect), will be unclear if the sample size is too small or trivial otherwise, but will never become positive or negative.

3 Provides a rigorous and principled approach to infer practical significance, and provides a rigorous distinction between practical and mechanistic significance. Inference addresses benefit and harm, and by doing so facilitates decision-making on the practical relevance of an effect.

4 Provides a refined classification of inferences using descriptors of the probability (see Table 1) of each of three outcome ranges (positivity/benefit, triviality/negligibility and negativity/harm) that can be made. As a result of refined inference, practically and mechanistically worthwhile effects are more likely to be detected as substantial and therefore deemed publishable; therefore, researchers are more likely to be able to draw useful conclusions from the data and publication bias can be reduced.

5 Estimates of required sample size are based on practical significance or mechanistic significance and researcher-defined smallest important effect.

5. The application of magnitude-based inference in user research

In this section we illustrate the benefits of magnitude-based inference with examples from user research. We start by presenting and interpreting results from usability studies comparing product versions (Sauro and Lewis, 2012). We then use the

results of various further examples of applying magnitude-based inference in user research as a basis for deriving some general observations. Next, we fully present an example of sample size estimation for a between-subjects design. Finally, we discuss magnitude-based inference for research designs beyond simple between-subjects and within-subjects.

5.1. Comparing two versions of a product

In usability-testing, inferential statistics are most commonly used to compare two product versions in order to assess usability and make appropriate important decisions (Rubin, 1994). This analysis is frequently employed when an organisation is considering to upgrade to a new version of a product, an organisation is intending to purchase one of two products for the same purpose or a manufacturer or software developer is deciding whether to launch a new product version to replace an existing version. Either a within-subjects design (each test user employs all different designs being compared) or a between-subjects design (each test user employs one of the different designs) is used to compare two designs or products. Here, we present two examples based on cases published by Sauro and Lewis (2012).[3]

Example 1, between-subjects design. In a usability test of a customer-relation management (CRM) application, test users have to add a customer's contact details to the application (Sauro and Lewis, 2012, pp. 72-73). Eleven test users employ the existing version, nine the new Version A and another nine new Version B. A cost-benefit analysis has demonstrated that in order to earn back the cost of the new version a reduction of at least 20% in time-on-task is needed; therefore, this value is

[3] A further illustrative example is presented in Online Appendix A.

set as the smallest important effect, so with the new versions the task should take at least 20% less time than with the original version.

Mean (±*SD*) time-on-task was 37±23 s for the existing version, 18±13 s for new Version A and 22±11 s for new Version B.[4] The application of magnitude-based inferences enabled the research team to calculate the % chances of benefit (or harm) as a result of adopting the new versions, with reference to a change of 20%, which equates to 7 s. The following inferences are based on the results of *t* tests for independent samples, conducted with a common statistics package, which are subsequently used as input into a spreadsheet for magnitude-based inference (see Supplementary Material A [and *http://sssl-staffweb.tees.ac.uk/U0011128/mbi/*]; main results are presented in Table 2). This is the original sheet 'Confidence limits & clinical chances' (at *http://www.sportsci.org/resource/stats/*; Hopkins, 2007) that has been modified for illustrative purposes; we recommend that researchers use Hopkins' original sheet for their research and reference accordingly.

_____ Insert Table 2 about here. _____

_____ Interested readers consult Supplementary Material A about here. _____

Compared to the existing version, time-on-task with new Version A was 19 s faster (90%-confidence interval 4 to 34 s), with a 90.4% chance of benefit, a 9.3% chance of negligibility, 0.4% chance of harm and an odds ratio of benefit to harm of 2929 in relation to the smallest important effect of 7 s. The use of the new Version A over the existing version of CRM application is therefore likely to be beneficial, the chances of harm are less than 0.5% and the chances of benefit are greater than

[4] Sauro and Lewis (2012) analysed the mean difference between the existing version and new Version A with *t* test for independent samples and NHST rather than with magnitude-based inference.

25%. Given this result and provided the same pattern of results occurs with other essential tasks as the one studied here, new Version A is recommended (practical inference). Furthermore, the effect is likely to be positive and unlikely to be trivial (mechanistic inference).

Compared to the existing version, time-on-task with new Version B was 15 s faster (90%-confidence interval 1 to 30 s), with a 82.8% chance of benefit, a 16.6% chance of negligibility, 0.6% chance of harm and an odds ratio of benefit to harm of 749 in relation to the smallest important effect of 7 s. The effect of using the Version B over the existing version of the CRM application is therefore unlikely negligible and not recommended (practical inference), as the chances of harm are greater than 0.5% while the chances of benefit are greater than 25%.[5] Furthermore, the effect is likely to be positive and unlikely to be trivial (mechanistic inference).

Example 2, within-subjects design. In a usability test of an accounting application, test users have to create an expense report (Sauro and Lewis, 2012, pp. 72-73). Twenty-one test users employ the existing version, new Version X and new Version Y in counterbalanced order. As a smallest important effect size, a 20% time-saving was required for the new versions to make a demonstrable difference, according to a cost-benefit analysis.

Mean (± *SD*) time-on-task was 231±88 s for the existing version, 152±45 s for new Version X and 207±45 s for new Version Y.[6] The application of magnitude-based inferences enabled the research team to calculate the % chances of benefit (or

[5] Note that if in advance of data collection, the researcher had decided to use Hopkins et al.'s (2009) alternative decision rule (the odds ratio of benefit to harm must be greater than a cut-off value of 66 for an effect to be beneficial) then the inference would be that the effect is beneficial, as the obtained odds ratio exceeds the cut-off.

[6] Sauro and Lewis (2012) analysed the mean difference between the existing version and new Version X with *t* test for paired samples and NHST rather than with magnitude-based inference.

harm) as a result of adopting the new versions, with reference to a change of 20%, which equates to 58 s. The following inferences are based on the results of *t* tests for paired samples, conducted with a common statistics package, which are subsequently used as input into a spreadsheet for magnitude-based inference (see Supplementary Material A; main results are presented in Table 3).

_____ Insert Table 3 about here. _____

_____ Interested readers consult Supplementary Material A about here. _____

Compared to the existing version, time-on-task with new Version X was 80 s faster (90%-confidence interval 56 to 104 s), with a 93.4% chance of benefit, a 6.6% chance of negligibility, 0.0% chance of harm and an odds ratio of benefit to harm of 6757803893 in relation to the smallest important effect of 58 s. The use of new Version X over the existing version of CRM application is therefore likely to be beneficial, with chances of harm less than 0.5% and the chances of benefit greater than 25%. Given this result and provided the same pattern of results occurs with other essential tasks as the one studied here, the adoption of new Version X is recommended (practical inference). Furthermore, the effect is likely to be positive and unlikely to be trivial (mechanistic inference).

Compared to the existing version, time-on-task with new Version Y was 25 s faster (90%-confidence interval 1 to 49 s), with a 1.4% chance of benefit, a 98.6% chance of negligibility, 0.0% chance of harm and an odds ratio of benefit to harm of 3579 in relation to the smallest important effect of 58 s. The effect of using the new version over the existing version of CRM application is therefore unclear and not recommended (practical inference), as the chances of benefit are less than 25%

even though the chances of harm are less than 0.5%.[7] Furthermore, the effect is very likely to be trivial (mechanistic inference). Following these two examples, a further illustrative example is presented in Online Appendix A.

_____ Interested readers consult Online Appendix A about here. _____

5.2. General observations

We make the following general observations from the results and inferences presented in a further example of a two-group between-subjects design in Online Appendix B and (mean-BS-SUS-1 in) Supplementary Material A.

1 The results of both practical and mechanistic inference for the same data set differ as a function of researcher-defined cut-offs for clinically or practically important harmful/negative and beneficial/positive effect sizes.

2 A wide range of practical and mechanistic inferences is observed. These types of inference are refined in the sense that they provide a considerable differentiation of clear results ('is almost certainly (not) harmful/trivial/beneficial', 'is very (un)likely to be harmful/trivial/beneficial', 'is (un)likely to be harmful/trivial/beneficial', 'is probably (not) harmful/trivial/beneficial', 'is possibly (not) harmful/trivial/beneficial'/may (not) be harmful/trivial/beneficial').

3 Practical and mechanistic inference mostly concur. However, an unclear effect according to mechanistic inference ('unclear; get more data') may be clear in practical inference (e.g. 'possibly harmful, unlikely beneficial; don't use' or 'unlikely

[7] Note that if in advance of data collection, the researcher had decided to use Hopkins et al.'s (2009) alternative decision rule (the odds ratio of benefit to harm must be greater than a cut-off value of 66 for an effect to be beneficial) then the inference would be that the effect is beneficial, as the obtained odds ratio exceeds the cut-off. In this case, the odds-ratio rule would produce a surprising result. Such surprising results could be avoided by the following perhaps more acceptable odd-ratio rule: odds ratio greater than 66 and chances of benefit greater than 25%.

harmful, unlikely beneficial; don't use'). The reverse may also occur with a clear result in mechanistic inference (e.g., 'possibly trivial') that is unclear according to practical inference ('unclear; don't use; get more data'). As explained in Section 4.2, such differences in results between these two types of inferences cannot be avoided. These differences should not necessarily be seen as a problem, but as an opportunity to clarify the difference between practical and mechanistic importance. There is a time-honoured idea that a particular set of results may be theoretically interesting, but practically unimportant. The approach of magnitude-based inference enables, or rather forces, researchers to make explicit their choices of thresholds for mechanistically (theoretically) important and practically important effects. As a consequence, an inference of practical importance, together with an inference of a lack of mechanistic importance from the same data can be defended by a researcher who has made these choices. The results presented in Supplementary Material A may even overestimate the concordance between the conclusions from mechanistic inference and practical inference. This is because in these results the same thresholds are chosen for practically beneficial and mechanistically positive effects on the one hand and for practically harmful and mechanistically negative effects on the other hand. However, depending on the seriousness of a harmful effect, a researcher may, for example choose a higher threshold for practical significance (e.g., $d = 0.8$) and a lower threshold for mechanistic importance (e.g., $d = 0.2$), leading to different conclusions from practical inference (more conservative here) and mechanistic inference (less conservative here).

4 The results of practical and mechanistic inference concur about half of the time with those of statistical inference. When the results differ, the latter is more conservative.

5.3. Sample size estimation

Table 4 presents condensed results for sample size estimation of a two-group between subjects design. (Full results are presented in Supplementary Material B [and *http://sssl-staffweb.tees.ac.uk/U0011128/mbi/*], tabulated Sheet mean-t-test-BS. This is the original sheet 'Sample-size estimation' (at *http://www.sportsci.org/resource/stats/*; Hopkins, 2006a) that has been modified for illustrative purposes; we recommend that researchers use Hopkins' original sheet for their research and reference accordingly.) Required sample size decreases rapidly with increasing smallest important effect size in all three types of inference. Sample sizes for mechanistic and practical inference concur with smaller effect sizes, but with larger effect sizes mechanistic inference tends to require smaller sample sizes than practical inference does.

_____ Insert Table 4 about here. _____

_____ Interested readers consult Supplementary Material B about here. _____

5.4. Further designs

Supplementary Materials A and B present further examples, with data adapted from Sauro and Lewis (2012). The classification of designs is based on, and the examples themselves are adapted from this recent text on the use of statistical methods in user research. We encourage interested readers to explore those examples that appeal most to them and thereby develop a further appreciation for the range of applications of magnitude-based inference. However, to save valuable space and avoid repetition, we do not discuss the results of these examples. In any case, the conclusions are consistent with, and provide further support for those presented in Section 5.1.2. In Supplementary Material A, the first is a further

example of a comparison of means with two systems on time-on-task in a between-subjects design. Two examples of comparisons of means between two systems with within-subjects design and perceived-usability data and time data follow. Then there are two examples of comparison of a mean against a target (perceived-usability data and time-on-task data). Finally, examples of comparing proportions between two systems (between subjects, purchase ['conversion-rate'] data; within subjects, task-completion data) and comparing a proportion with a target (task-success data) are presented.[8] Supplementary Material B presents further examples of sample size estimation, for a within-subjects design comparing means and for a design comparing two proportions (between subjects, within subjects) or a proportion against a target.[9] The conclusions that the reader can draw from these examples concur with those drawn in Section 5.2.

In Section 5 we have illustrated magnitude-based inference with examples using basic research designs to further facilitate the reader's understanding of this type of inference. However, more complex research designs can also be analysed (e.g., Hopkins, 2006b). Currently, the available spreadsheets (*http://www.sportsci.org/*) support the various analyses, including the following: raw difference between means and other *t*-distributed effect statistics; percent and factor difference between means and other log *t*-distributed effect statistics; rate ratio and other log-normally distributed effect statistics; correlation coefficient; standard deviation, coefficient of variation or root mean square error; ratio of two independent standard deviations; combining outcomes for subject groups or statistics. Analysis of the following types

[8] These examples use adjustments in calculating probabilities and sample sizes, as advocated by Sauro and Lewis (2012). Calculations are then made using the standard normal distribution as an approximation for the binomial distribution, as sample size is sufficient (Clark-Carter, 2009).

[9] Again, the standard normal distribution is used to provide approximations.

of design is supported: (1) basic experimental between-subjects designs and within-subjects designs with an independent variable having two levels; (2) complex multifactor experimental designs with two levels per factor – this is possible because both main effects and interaction effects can be tested directionally, with *df* = 1; (3) pre-post parallel-groups controlled designs with adjustment for a predictor, post-measure-only crossover design with adjustment for a predictor, and pre-post crossover design with adjustment for a predictor; (4) single-group designs (e.g., testing the sample mean against the known value of a population mean) and correlational designs.

Magnitude-based inference does not support analyses of complex designs with effects having more than one degree of freedom (e.g., as in analysis of variance). However, it has been proposed that omnibus tests (with *df* > 1) should not be conducted in any case and that, instead, (all) inferential statistics should be conducted through (planned, directional) contrasts (Rosenthal et al., 2000; see also Dienes, 2011), and these are supported by the spreadsheets. This will require researchers to think carefully about how to answer their research questions through the analysis of specific contrasts rather than relying on omnibus tests.

The fundamental reason is that magnitude-based inference, as described in this paper, cannot be applied when the tested effect is not directional, for example an omnibus effect in analysis of variance (ANOVA) or a model test of R^2 in multiple regression analysis. This is because unsigned (omnibus) effects are tested in these analyses, so it is logically impossible to test whether these effects are positive or negative.

For those who still wish to test an effect with *df* > 1, in the spirit of magnitude-based inference, it would be natural to conduct a minimum-effect test (Murphy & Myors, 1999), with the output of null-hypothesis-testing (e.g., *F*-value) as input. The smallest important effect in magnitude-based inference corresponds with what is called the researcher-defined minimal effect in minimum-effect tests. This effect can be tested either as a statistical-significance test (as in Murphy & Myors, 1999) or the same qualitative descriptors that are used in magnitude-based inference may be employed here (most likely, very likely etc.; see Table 1). If an omnibus minimum-effect test in ANOVA shows a substantial effect, then this could be followed up with specific directional comparisons by way of magnitude-based inference.

6. Other approaches to statistical inference

Even though the use of NHST is common in user research and elsewhere, several limitations of have been noted in the behavioural-research literature, with cogent calls for abandoning NHST altogether (e.g., Cohen, 1994; Cumming, 2014; Dienes, 2011; Rozeboom, 1997). From the perspective of magnitude-based inference, it is most important to note that (a) NHST does not require the researcher to specify a smallest important effect and (b) NHST does not use such an effect in making inferences. However, as shown in this paper, the results of NHST can be used as input for magnitude-based inference, addressing both of these issues. Apart from magnitude-based inference, other alternatives to NHST have also been proposed. In human-computer interaction, Landauer (1997) has suggested using the odds of obtaining the actual difference between two experimental conditions “merely by chance”, (1 - *p*-value)/*p*-value (p. 222). The counter-null statistic (Rosenthal & Rubin, 1994) is the non-null-magnitude effect size that has the same *p*-value as the null value of the effect size. It can be used to highlight the possibility of a (substantial)

effect when, in fact, the null hypothesis is not rejected. In that sense, it is reminiscent of findings in our examples presented in Online Appendix B showing that mechanistic or practical inference can demonstrate an important effect that goes undetected in statistical inference. p_{rep} (Killeen, 2005) is the probability of future research replicating the obtained result. *p*-intervals (Cumming, 2008) provide a range of likely *p*-values in the analysis of a particular data set. Each of these alternatives remains based on NHST, with inherent limitations (e.g., Cohen, 1994; Cumming, 2014; Dienes, 2011; Rozeboom, 1997). The same goes for the analysis of mixed-effect models, which has been proposed for designs with repeated measures, multi-level designs and designs with missing data.

Minimum-effect tests (Murphy & Myors, 1999) have been proposed as an advance over NHST. This is because the null hypothesis of no effect (the 'nil hypothesis') is replaced with a null hypothesis of a minimal effect (similar to the smallest positive/beneficial or negative/positive important effect in magnitude-based inference), specified by the researcher. Nonetheless, this approach does not make a distinction between practical and mechanistic inference and does not provide a refined classification of inferences as magnitude-based inference does. Minimum-effect tests share with magnitude-based inference the advantage that the test result is not an artefact of sample size, but is limited to retaining or rejecting the hypothesis of a minimal effect, does not distinguish positive/beneficial and negative/harmful effects, and does not distinguish between an unclear effect and a clear effect.

The use of confidence intervals has also been proposed as an alternative to NHST. For example, Cumming (2014) presents six approaches to interpreting confidence intervals, but does not use a smallest important effect size as a meaningful reference and does not provide guidance for making inferences in relation to such an effect

size. Furthermore, confidence intervals are used in equivalence testing and non-inferiority testing.[10] The purpose of the first is to establish that two treatments equally good, for example with respect to a specified size of mean difference delta, by verifying that the confidence intervals exceed a difference of neither -1/2delta nor + 1/2delta (Head et al., 2012). The purpose of the second is to establish that one treatment is not worse than a second treatment, for example with respect to a specified size of mean difference delta, by verifying that the confidence intervals do not exceed a difference of -delta (Head et al., 2012). First, the statistics that are calculated, and second, arguably even more important, the types of conclusion that can be drawn from equivalence testing differ from those that apply to magnitude-based inference. In particular, when used in conjunction with NHST, equivalence testing allows a limited range of potential conclusions: equivalence (which could be considered to mean the same as 'trivial' in magnitude-based inference), difference (reject the null hypothesis) or otherwise indeterminate (Tryon, 2001). In contrast, magnitude-based inference allows a considerably richer set of potential conclusions, consisting of unclear (collect more data) and otherwise combinations of the form <Qualifier 1> beneficial/<Qualifier 2> trivial/<Qualifier 3> harmful, where each qualifier is from the range of seven descriptors presented in Table 1.

Bayesian statistical inference has also been proposed as an alternative to NHST and shares with magnitude-based inference the advantage that inference is not an artefact of sample size (Kruschke, 2011). In this approach the researcher's prior belief is adjusted based on the data that have been collected. However, providing believable estimates of prior beliefs is considered a major obstacle (Bland & Altman,

[10] In contrast, superiority testing aims to show that one treatment is better than a second treatment. This is typically done through NHST, but – as argued theoretically and demonstrated empirically in this paper – magnitude-based inference provides an attractive alternative.

1998; see also Dienes, 2011). Moreover, there is doubt about the accessibility, comprehensibility and usability of this approach for researchers (Hopkins, 2006a). For instance, although it is possible to specify a range of effect sizes corresponding with the experimental hypothesis through a distribution of prior probabilities on effect sizes, the understandability and usability, for (user) researchers, of this specification in relation to the research design of an experiment or other study remains in doubt. In contrast, effect size measures, such as *d* for difference in mean score between a control condition and an experimental condition, that are used in magnitude-based inference to define the smallest important effect are widely published and have a familiar interpretation. In any case, in Bayesian inference, as in NHST and magnitude-based inference, the obtained effect size needs to be reported as well (Wetzels et al., 2011).

Still, Bayesian inference through model comparison (Kruschke, 2011), for example by way of the Bayesian *t* test (Rouder et al., 2009; Dienes, 2011) has the advantage of avoiding a dichotomous (as in NHST) or trichotomous (as in equivalence testing, combined with NHST) decision. The Bayes factor is an odds that is used in the Bayesian *t* test and represents the change in a researcher's belief towards the experimental hypothesis as a result of the collected data (Bayes factor > 3) or towards the null hypothesis (Bayes factor < 1/3). It can be interpreted through a refined scheme for inference (Jeffreys, 1961, cited in Wagenmakers et al., 2011), ranging from extreme/very strong/strong/substantial/anecdotal evidence for the null hypothesis to no evidence to anecdotal/substantial/strong/very strong/extreme evidence for the experimental hypothesis. However, it is important to note that the Bayes factor "can vary dramatically depending on the choice of alternative-model prior distribution" (Kruschke, 2011, p. 307). Moreover, a potential disadvantage of

using the Bayes factor is that its use may suffer from the same problem of using the odds of benefit to harm in magnitude-based inference (see our example in Section 5.1.1): even if the (absolute value of the) probability supporting the favoured hypothesis is small, the odds may be large, thereby providing inflated evidence in favour.

As an alternative to Bayesian model comparison, Kruschke (2011) offers Bayesian interval parameter estimation. This involves calculating the 95% highest-density interval of parameter values for each model parameter, containing 95% of the parameter's distribution. The values contained in this interval have higher credibility than those outside of it. This first interval is compared with a second interval around the null value (region of practical equivalence) to evaluate the credibility of this value. The null value can be rejected (no overlap between the two intervals) or accepted (the first interval is contained in the second interval) or judgement can be suspended (partial overlap). It is important to note that the proportion of the first interval contained in the second, which forms the basis of inference here, can be highly sensitive to the limits of the first interval (Kruschke, 2011).

Recently, Trafimow and Marks (2015) have – as editorial policy – declared an outright ban on the use of NHST and confidence intervals. They also find much of Bayesian inference problematic. Instead, they demand the reporting of "strong descriptive statistics, including effect sizes" (p. 1).

There has been a growth in the application of magnitude-based inference in meta-analyses in sport and exercise science (e.g., Weston et al., 2014). Meta-analyses are essential tools for summarising evidence accurately; statistical methods are used to summarise and combine the results of independent studies (Liberati et al., 2009).

Analysis of the overall effect – the main output statistic of meta-analysis – via magnitude-based inference again provides a more informative qualitative descriptor for the 'pooled effect'. For example, an overall meta-analysed effect size (standardised mean difference) of 1.0 would be qualified as moderate, yet using magnitude-based inference such an effect could be reported as possibly/likely/very likely/almost certainly moderate, depending upon the width of the associated confidence interval (Weston et al., 2014). Further, when an unclear effect is present, qualitative descriptors can be provided for the upper and lower end of the confidence interval to provide the likely range of the effect. For example, in the meta-analysis of Weston et al. (2014) an unclear effect was reported for the effect of high-intensity training on power output, yet the authors quantified the upper and lower end of the likely range of effects and reported that: "training had an *unclear* effect on sprint power that could at most be a *moderate beneficial* or a *small harmful* effect".

7. Limitations

We have argued for magnitude-based inference, with important benefits for making inferences in user research. However, limitations of this type of inference can be identified, some apparent and others substantive.

7.1. Apparent limitations

In magnitude-based inference, the researcher needs to make several choices or accept recommended choices, for example about Type-1 and Type-2 practical-error rates, the smallest important effect and the mapping of quantitative probabilities onto qualitative descriptors. Recommended choices are cogently argued (e.g., Hopkins, 2007) and may be reasonable for many user research studies. However, it is important to note that magnitude-based inference forces researchers to make these

choices explicit, assisted (among other things) by a language to describe probabilities qualitatively (see Table 1). A researcher can choose their own mapping, but would have to justify their choices to convince their reviewers/editors/readers. Instead of presenting an apparent limitation, this need to make choices explicit highlights limitations of NHST, in which Type-I and Type-II error rates are typically constrained by conventions that are deemed too conservative (leading to conclusions of unclear results that could be important and sample sizes that may be larger than necessary; Hopkins, 2007) and in which the definition of a smallest important effect has no influence on the inference that is made. By offering a rich range of possible inferences (aided by a mapping of probability ranges onto descriptors) rather than the two or three inferences in NHST, magnitude-based inference avoids the temptation of 'fiddling' the significance level that is used in NHST.

By definition, the choice of smallest important effect size, which then affects inferences, remains subjective. However, an informed choice can be guided by considerations discussed in Section 4.1.3 and the unrealistic choice of an effect size of 0 in NHST is also subjective. Still, a concern might be that an unscrupulous researcher might 'revise' the choice of smallest important effect to produce a more favourable inference. However, an attempt to do so would be futile. This is because researchers using MBI will have to convince their reviewers/editors/readers that the researchers' choice of smallest important effect is sensible. Reviewers/editors/readers will not be convinced if the researchers' choice is not presented with a credible rationale. In contrast, with NHST researchers select a null effect without any attempt to convince the reader that is a sensible choice, even

though many publications have previously argued before that this choice is not sensible.

The choice of a common smallest important effect size facilitates the comparison of magnitude-based inferences between studies. However, even when different choices were originally made by different researchers conducting different studies, the results can still be compared by 'equalising' the choices, if the results are presented in one of the various publicly available spreadsheets (*http://www.sportsci.org/*). For example, the same set of studies can be compared under a small, medium or large smallest important effect size.

Although the publicly available spreadsheets support various common effect statistics, it cannot be claimed that any conceivably useful statistic will be supported. Nevertheless, existing spreadsheets (*http://www.sportsci.org/*) offer a range of analyses. In any case, based on existing statistical theory (Smithson, 2003), spreadsheet calculations can be added to support other directional statistics, thus making this type of inference even more practically useful.

A requirement for the analyses presented in this paper is normality of the sampling distribution of the outcome statistic. Hopkins et al. (2009, p. 8) take the view that "the central-limit theorem ensures that the sampling distribution is close enough to normal for accurate inferences, even when sample sizes are small (about 10) and especially after a transformation that reduces any marked skewness in the dependent variable or nonuniformity of error". They discourage the use of non-parametric analyses because of a lack of power, and the inability to adjust for covariates and to permit inferences about magnitude. However, bootstrapping is offered (with a supporting spreadsheet) as an alternative in cases where a normal

sampling distribution cannot be assumed or established, for instance when non-linear relationships are modelled (Hopkins, 2012), to provide robust estimates of confidence intervals and magnitude-based probabilities.[11] Moreover, bootstrapping can be used more generally for any effect statistic by researchers who take a more sceptical view of the normality of sampling distributions.

7.2. Substantive limitations

Welsh and Knight (2014) analyse the use of magnitude-based inference in the comparison of two means. Their substantive claim against magnitude-based inference is that it inflates the probability of detecting an effect that does not exist. Batterham and Hopkins (2014) provide evidence against this claim and other claims regarding magnitude-based inference made by Welsh and Knight (2014). Irrespective of the veracity of the claim of inflated false alarms in magnitude-based inference, researchers who are concerned about this could reduce this possibility. This can be achieved by prudently deciding not to make inferences of a substantial effect of positivity/benefit with outcomes that are qualified as possibly positive/beneficial (chances between 25% and 75%), but to make inferences of a substantial effect only with higher chances (at or above 75%).

The decision rules themselves for mechanistic inference and practical inference are cogently argued (Batterham & Hopkins, 2006; Hopkins, 2007), but do not necessarily take all relevant factors into account (irrespective of whether one agrees on the recommended values for, for example, the numerical values of the probability for a

[11] Bootstrapping is a statistical resampling technique that is used to create confidence intervals and test statistical hypotheses, without the need to make distributional assumptions. From the original sample data, a large number of samples (e.g., $N = 5000$) is taken by sampling with replacement. From each sample the statistic that is of interest (e.g., the mean) is calculated and used to create a confidence interval or an inferential statistic to test a hypothesis.

Type-1 practical error and a Type-2 practical error). For example, as presented in the literature, magnitude-based inference does not (yet) take into account the (financial) value inputs to and outputs from using a harmful or beneficial effect (Hopkins, 2007). A practical decision may have to be based on these considerations as well. Existing literature on cost-justifying usability work (e.g., Bias & Mayhew, 2005) does not consider the possibility of using magnitude-based inference in user-research studies. A related approach would be to use utility analysis to define a smallest important effect (Murphy & Myors, 1999). This takes into account the financial value of the benefit (effect/dependent variable) and cost (for example change from a [cheaper] existing to new [more expensive] version of product or level of an independent variable). The idea is that, if sufficient variance in the dependent variable is explained by the independent variable to offset the cost, then the change to a new product can be justified. A utility equation is solved to determine the minimum effect size, expressed as a correlation between the independent and the dependent variable, to achieve this offset. Other variables in the equation are the standard deviation of the dependent variable and the cost (e.g., price increase of new product over existing product).[12] In case a different effect size measure (e.g., the standardised mean difference *d*) than the correlation coefficient is desired, the value of the smallest important correlation value can be converted (Clark-Carter, 2009) to the desired effect size measure, once the equation has been solved.

Still, even if the decision rules of magnitude-based inference are not perfect, as theoretically argued in this paper and illustrated with examples of the results of analysing user-study data, decision-making in magnitude-based inference can be

[12] $\Delta U = r_{xy} \times SD_y - C$, projected overall benefit ΔU, correlation coefficient r_{xy}, standard deviation of the dependent variable SD_y, cost C.

considered as an important advance over decision-making in NHST. In particular, this inference can help to avoid making decisions that are based on the unrealistic smallest important null-effect size and that may be too conservative, by taking into account the smallest important effect in the inference that is being made and by using meaningful and intuitively appealing descriptors for qualifying probabilities, and inferences are never an artefact of sample size.

8. Recommendations

Although the use of magnitude-based inference has apparently not yet been reported in the literature on user research and human-computer interaction more generally, the generic recommendations for the application of this approach (see below) are in principle the same as those in sport and exercise science. However, in all disciplines, researchers will need to choose a smallest important effect size, which will differ, depending on specific considerations in the design of a particular experiment or other study. The supporting spreadsheets are flexible in the sense that they allow researchers to specify various parameters, including the smallest important effect size. Moreover, a range of analyses is available.

In this paper, we have presented various benefits that magnitude-based inference offers and provided illustrations from user research. Naturally, it is up to our readers to decide which approach to use in their research, taking into account our argument and results. Our recommendations are as follows when magnitude-based inference is used.

1 Where possible, plan sample size using magnitude-based inference. Plan sample size in advance or, better according to Hopkins (2006a), ‘on the fly’ (in a group sequential design, where data are collected until a clear outcome emerges).

Otherwise, retrospectively compare the actual sample size with the required sample size to gain an understanding of the appropriateness of the former.

2 Consistent with existing practice, analyse data using tests of the null hypothesis with a common statistical package or spreadsheet. However, augment existing practice by making more informative use of the results as input for magnitude-based inference.

3 Always analyse data using mechanistic inference. Also use practical inference for effects that have a direct practical application.

4 Use appropriate spreadsheets to facilitate sample size estimation and magnitude-based inference (*http://www.sportsci.org/*).

5 When preparing for journal publication, write to the journal editor arguing cogently why it is appropriate to use magnitude-based inference in your research. In your section Data Analysis explain the specific magnitude-based inference that you have used (e.g., Barnes et al., 2015).

9. Conclusion

The *aim* of this paper is to be translational by theoretically making the case for an alternative approach to statistical data analysis, called magnitude-based inference, with several benefits for research in human-computer interaction, and by empirically illustrating this approach and its advantages with examples from user research. We discussed theoretical benefits of the approach and our example numerical results from designs that are common in user research showed in detail how the approach can be advantageous. Researchers can continue their existing practice of analysing data using tests of the null hypothesis, but make more informative use of the results

by using these as input for magnitude-based inference making. In conclusion, magnitude-based inference is an attractive alternative approach, which can produce informative inferences in user research. We look forward to the fruitful application of this approach in user research, human-computer interaction more generally and beyond.

Acknowledgements

The authors are grateful to Ioannis Vasilopoulos, Gabor Aranyi and Debora Jeske for helpful comments on earlier versions of this work. The authors also thank the handling editor Paul Mulholland and three anonymous reviewers.

References

Atzeni, A., Su, T., Baltatu, M., D'Alessandro, R., Pessiva, G., 2014. How dangerous is your android app? An evaluation methodology. MobiQuitous - Int. Conf. Mob. Ubiquitous Syst.: Comput., Netw. Serv., 130-139.

Barnes, K.R., Hopkins, W.G., McGuigan, M.R., Kilding, A.E., 2015. Warm-up with a weighted vest improves running performance via leg stiffness and running economy. J. Sci. Med. Sport 18, 103-108.

Batterham, A.M., Hopkins, W.G., 2014. The case for magnitude-based inference. Med. Sci. Sports Exerc. 47, 885.

Batterham, A.M., Hopkins, W.G., 2006. Making meaningful inferences about magnitudes. Int J Sports Physiol Perform 1, 50-57.

Bias, R., Mayhew, D., 2005. Cost-justifying usability.

Bland, J.M., Altman, D.G., 1998. Bayesians and frequentists. BMJ 317, 1151-1160.

Cairns, P., Cox, A., 2008. Research Methods for Human-Computer Interaction. Cambridge University Press, Cambridge.

Clark-Carter, D., 2009. Quantitative Psychological Research, 3rd ed. Psychology Press, Hove.

Cohen, J., 1994. The earth is round ($p < .05$). Am. Psychol. 49, 997-1003.

Cumming, G., 2008. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science 3, 286-300.

Cumming, G., 2014. The New Statistics: Why and How. Psychol. Sci. 25, 7-29.

De Oliveira, R., Cherubini, M., Oliver, N., 2013. Influence of personality on satisfaction with mobile phone services. ACM Trans. Comput. -Hum. Interact. 20.

Dienes, Z., 2011. Bayesian versus orthodox statistics: Which side are you on? Perspect. Psychol. Sci. 6, 274-290.

Drury, C.G., 2015. Human factors/ergonomics implications of big data analytics: Chartered Institute of Ergonomics and Human Factors annual lecture. Ergonomics 58, 659-673.

Grossman, T., Fitzmaurice, G., Attar, R., 2009. A survey of software learnability: Metrics, methodologies and guidelines. Conf Hum Fact Comput Syst Proc, 649-658.

Head, S.J., Kaul, S., Bogers, A.J.J.C., Kappetein, A.P., 2012. Non-inferiority study design: Lessons to be learned from cardiovascular trials. Eur. Heart J. 33, 1318-1324.

Hopkins, W.G., 2006a. Estimating sample size for magnitude-based inference. Sport Science 10, 63-70.

Hopkins, W.G., 2006b. Spreadsheets for analysis of controlled trials, with adjustment for a subject characteristic. Sport Science 10, 46-50.

Hopkins, W.G., 2007. A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a p-value. Sport Science 11, 16-20.

Hopkins, W.G., 2012. Bootstrapping inferential statistics with a spreadsheet. Sport Science 16, 12-15.

Hopkins, W.G., Marshall, S.W., Batterham, A.M., Hanin, J., 2009. Progressive statistics for studies in sports medicine and exercise science. Med. Sci. Sports Exerc. 41, 3-12.

Hornbæk, K., 2013. Some whys and hows of experiments in human-computer interaction. Foundations and Trends in Human–Computer Interaction 5, 299-373.

Hornbæk, K., Sander, S.S., Bargas-Avila, J., Simonsen, J.G., 2014. Is once enough? on the extent and content of replications in human-computer interaction. Conf Hum Fact Comput Syst Proc, 3523-3532.

International Standards Organization (ISO), 1998. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) - Part 11: Guidance on Usability. Author, Geneva.

Kaptein, M., Robertson, J., 2012. Rethinking statistical analysis methods for CHI. Conf Hum Fact Comput Syst Proc, 1105-1113.

Killeen, P.R., 2005. An alternative to null-hypothesis significance tests. Psychol. Sci. 16, 345-353.

Komarov, S., Reinecke, K., Gajos, K.Z., 2013. Crowdsourcing performance evaluations of user interfaces. Conf Hum Fact Comput Syst Proc, 207-216.

Kruschke, J.K., 2011. Bayesian assessment of null values via parameter estimation and model comparison. Perspect. Psychol. Sci. 6, 299-312.

Landauer, T.K., 1997. Behavioral research methods in human-computer interaction, in Helander, M.G., Landauer, T.K., Prabhu, P.V. (Eds.), Handbook of Human-Computer Interaction, 2 ed. Elsevier, Amsterdam, pp. 203-227.

Lazar, J., Feng, J.H., Hochheiser, H., 2010. Research Methods in Human-Computer Interaction. Wiley, Hoboken, NJ.

Lenth, R.V., 2006-2009. Java Applets for Power and Sample Size [Computer software]. .

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. Ann. Intern. Med. 151, W-65-W-94.

Murphy, K.R., Myors, B., 1999. Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. J. Appl. Psychol. 84, 234-248.

Nielsen, J., 1993. Usability Engineering. Academic Press, Boston, MA.

Oladimeji, P., Thimbleby, H., Cox, A.L., 2013. A performance review of number entry interfaces. Lect. Notes Comput. Sci. 8117 LNCS, 365-382.

Pagallo, U., 2013. Robots in the cloud with privacy: A new threat to data protection? Comput Law Secur. Rev. 29, 501-508.

Pinker, S., 2014. The Sense of Style. Allen Lane, London.

Purchase, H., 2012. Experimental Human-Computer Interaction: A Practical Guide with Visual Examples, Cambridge University Press ed. Cambridge.

Raij, A., Ghosh, A., Kumar, S., Srivastava, M., 2011. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. Conf Hum Fact Comput Syst Proc, 11-20.

Rosenthal, R., Rubin, D.B., Rosnow, R.L., 2000. Contrasts and Effect Sizes in Behavioral Research. Cambridge University Press, Cambridge, UK.

Rosenthal, R., Rubin, D.B., 1994. The counternull value of an effect size: a new statistic. Psychological Science 5, 329-334.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G., 2009. Bayesian t tests for accepting and rejecting the null hypothesis. Psychonom. Bull. Rev. 16, 225-237.

Rozeboom, W.W., 1997. Good science is abductive, not hypothetico-deductive, in Harlow, L.L., Mulaik, S.A., Steiger, J.H. (Eds.), What if there were no Significance Tests?. Erlbaum, Mahwah, NJ, pp. 335-392.

Rubin, J., 1994. Handbook of Usability Testing. Wiley, New York.

Sauro, J., 2011. A Practical Guide to the System Usability Scale. Create Space, Denver, CO.

Sauro, J., Lewis, J.R., 2012. Quantifying the User Experience. Morgan Kaufmann, Amsterdam.

Schumacher, R., 2009. The Handbook of Global User Research. Morgan Kaufmann, Boston.

Scriven, M., 1967. The methodology of evaluation, in Tyler, R., Gagne, R., Scriven, M. (Eds.), Perspectives of Curriculum Evaluation. Rand McNally, Chicago, pp. 39-83.

Smithson, M., 2003. Confidence Intervals. Sage, Thousand Oaks, CA.

Stewart, A.M., Hopkins, W.G., 2000. Consistency of swimming performance within and between competitions. Med. Sci. Sports Exerc. 32, 997-1001.

Trafimow, D., Marks, M., 2015. Editorial. Basic Appl. Soc. Psychol. 37, 1-2.

Tryon, W.W., 2001. Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. Psychol. Methods 6, 371-386.

Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H.L.J., 2011. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). J. Pers. Soc. Psychol. 100, 426-432.

Welsh, A.H., Knight, E.J., 2014. "Magnitude-based Inference": A statistical review. Med. Sci. Sports Exerc. 47, 874-884.

Weston, M., Taylor, K.L., Batterham, A.M., Hopkins, W.G., 2014. Effects of low-volume high-intensity interval training (HIT) on fitness in adults: A meta-analysis of controlled and non-controlled trials. Sports Med. 44, 1005-1017.

Wetzels, R., Matzke, D., Lee, M.D., Rouder, J.N., Iverson, G.J., Wagenmakers, E.-J., 2011. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. Perspect. Psychol. Sci. 6, 291-298.

Wilkinson, L., 1999. Statistical methods in psychology journals: Guidelines and explanations. Am. Psychol. 54, 594-604.

Wiseman, S., Cox, A.L., Brumby, D.P., Gould, S.J.J., O'Carroll, S., 2013. Using checksums to detect number entry error. Conf Hum Fact Comput Syst Proc, 2403-2406.

Wixon, D., 2011. Measuring fun, trust, confidence, and other ethereal constructs: It isn't that hard. Interactions 18, 74-77.