Int. J. Human-Computer Studies 88 (2016) 38–50

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Magnitude-based inference and its application in user research ☆

Paul van Schaik *, Matthew Weston

School of Social Sciences, Business and Law, Teesside University, Middlesbrough TS1 3BA, United Kingdom

ARTICLE INFO

Article history:
Received 20 November 2014
Received in revised form
27 November 2015
Accepted 2 January 2016
Communicated by Dr. P. Mulholland
Available online 8 January 2016

Keywords:
User research
Quantification
Statistics
Inference
Usability testing
User-experience

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.

"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

☆This paper has been recommended for acceptance by Dr. P. Mulholland.

* Corresponding author. Tel.: +44 1642 342320; fax: +44 1642 342301.
E-mail address: p.van-schaik@tees.ac.uk (P. van Schaik).

http://dx.doi.org/10.1016/j.ijhcs.2016.01.002
1071-5819/

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 and 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 and 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, 1999; Cairns and Cox, 2008; Kaptein and Robertson, 2012; Purchase, 2012; Hornbæk, 2013) and elsewhere (Wilkinson, 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 (2006–2009) 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