Evaluating and contrasting Human-Computer Interactions in two studies, and the importance of user testing in website design

Aiden Baker-Gabb — u7436429

26 May 2024

Executive summary

This report succinctly evaluates two studies involving Human-Computer Interaction (HCI). It assesses the overall findings of both studies, their strengths and limitations, and further assesses potential strategies to improve results. This report contrasts the differences between the two studies and also highlights how they both handle HCIs. The final section of this report uses the two studies to establish that initial experimental expectations can often be altered or even reversed after receiving data. Using this premise, the importance of user-based experiments and testing in developing accessible, functional, and credible websites, is established.

Study I — Felt depression is different to algorithmic depression

Experiment one, namely, 'Felt depression is different to algorithmic depression: A user experiment using an image processing depression dataset', investigated the accuracy of different modes of predicting the correct 'depression status' in 29 videos of subjects from the AVEC 2014 database. The different modes included the conscious decision of each of the 10 participants, patterns in subconscious biological markers in the participants, and an image processing algorithm. Each video in the AVEC database contained a German-speaking subject, with a depression status of minimal, mild, moderate, or severe.

Participants in the study were equipped with a galvanic skin response (GSR) sensor — in order to measure skin conductance, as well as an ECG — to measure heart rate variability (HRV). Outputs from the GSR and ECG sensors, dubbed *primary features,* were used to derive a number of *secondary features*. These secondary features were used as input for an Artificial Neural Network (ANN). A Genetic Algorithm (GA) was implemented in order to select the most useful secondary features, and an Extreme Learning Machine was used to increase the training speed. This approach was termed GA-ANN-ELM.

The study found that the hybrid GA-ANN-ELM approach was the most successful at predicting the correct depression status of subjects, with a success rate of 79%. In contrast to the participants' subconscious physiological markers, their conscious decision yielded an accuracy of 31.44%. This was even less than the image processing algorithm that could successfully identify the status of subjects 55.2% of the time.

Strengths

1. This study elucidated the potential power of subconscious physiological markers paired with an artificial neural network, in assessing the psychology of others.
2. This study showed that an image processing algorithm can more accurately determine the psychological state of subjects than the humans in the study.

Limitations

1. The study gave a limited description of the AVEC videos that were used.
2. This study did not mention the patterns in a subject's GSR or HRV that were identified by the Genetic Algorithm to conclude that a subject was subconsciously identifying a certain depression status.
3. This study employed a limited selection of participants. More participants from different backgrounds and varying age groups could have improved or disproved the apparent findings of this study.

Study II — Popular eReaders

Experiment two, namely, 'Popular eReaders', evaluated the user interface and intuitive design of four different eReaders, Nook, Kobo, Kindle, and Sony. The study recognised the growing popularity of eReaders, and used a scenario-based evaluation approach, in contrast to the more popular price comparison approach.

This study emulated a scenario in which a participant had just been given an eReader, and in order to use it, they had to find a certain point of information within the device. Participants worked in groups of two, where the first would attempt to complete the assigned task on a selected eReader, and the second would act as a scribe, writing down feedback from the first participant. The two participants would then swap roles for the next eReader.

The study had minimal results and only produced statistically significant data in the evaluation that participants gave for one task. This task involved reporting information from a paragraph within a specific section of a document. The participants favoured the Kobo the most, while they rated the Nook as the lowest, on average.

Strengths

1. The study highlights the importance of iterating over the design of the user interface, with large-scale user testing and feedback, amongst a wide variety of age groups, in order to create an intuitive design that caters to every customer.

Limitations

1. This study recognised, but did not proceed with further investigation, into the previous experience participants had had with eReaders, and how this may have impacted their results.

2. The study employed only a small selection of students from a very specific background to participate. This limited the study's ability to compare the evaluations each of the pairs gave, to the average consumer. The study would have yielded more generally applicable findings if it had employed participants from a broader range of age groups and backgrounds. For example, computer science students all around 25 years old, are likely to have fantastic technological intuition and experience, while 60-year-olds and above may have much less intuition or experience with such technology, and thus approach the evaluation from a completely different perspective. Examining the overall age distribution of eReader users would have also played an important role in drawing general conclusions from the results.
3. The paired structure of participants may have significantly influenced the evaluations given for each eReader. Evaluators (P1) were forced to relay their evaluation to a scribe (P2), who then recorded the evaluator's rating for different aspects of the eReader. The rating that P1 gave to P2 may be heavily influenced by what P1 thinks P2 expects P1 to say. These expectations that P1 places on their feedback to P2 may be pressured more or less depending on P1 and P2's relationship. Furthermore, the evaluation that P1 gives to P2 is subject to misinterpretation by P2. The information provided from P1 to P2 may also affect P2 in the same way, when it comes time for both participants to swap roles.
4. Poor method design. The method took the following approach: Each participant rated no more than one eReader. If participants were labelled P1, P2, … P11, P12, where each set of 2 participants starting from P1, and P2, all the way to P11, and P12, were a pair. Each member of each pair of participants took one turn at evaluating an eReader, and one turn at scribing the other participant's evaluation in the provided form. This method led every participant to personally evaluate no more than one eReader each. This study aimed to compare different eReaders based on participant evaluations but failed to let participants scale their ratings relative to their ratings for other eReaders, as they only got to evaluate one eReader. This method likely contributed to the lack of statistically significant results, as outlined by the authors. Better results would have likely been achieved if each participant had to evaluate all 4 eReaders.

Comparison of experiments

In both studies, participants were involved in Human Computer Interactions (HCIs). The nature of these HCIs differs substantially between the two studies. *Study I* deals with a very specific application of HCI. This study focuses on the potential outcomes of pairing ANNs with physiological markers in subjects in order to detect depression, and how this compares to visual-based processing algorithms. In this study, the HCI component manifests in the use of GSR and HRV sensors to detect subconscious depression-indicating patterns in a participant's physiology. Conversely, *Study II* demonstrates a less abstracted HCI and more broadly investigates the features of successful intuitive design among different brands of eReaders. This study more literally assesses the direct interaction between the human (participant) and the computer (eReader).

Study I executed a well-structured method, and drew succinct conclusions from the results. It successfully demonstrated the power of subconscious physiological markers in identifying the depression status of other humans, and the power of Artificial Neural Networks. By comparison, *Study II* was less scrutinous, and less structured, which may have contributed to the lack of statistical significance in the results, as noted by the authors. Both studies had very small sample sizes of participants, and limited diversity within these samples. This likely impacted the results and thereby conclusions of both experiments.

Conclusion

In web development and design, it is necessary to engage with users, giving them opportunities to provide meaningful feedback. Different age groups and educational backgrounds have different expectations of how a website should function. To produce an intuitive design that caters to a variety of different users, user testing is crucial.

Study II demonstrated the importance of intuitive design and the effect it can have on completing simple tasks. This is directly relevant in a web development and design setting, as the functionality and design of a website are critical to the user's experience, which directly influences potential sales or the credibility of a company.

While *Study II* does not generate a conclusion that is directly relevant to web design and development, it demonstrates that user participation is vital in the design process as initial expectations within an experimental model can be quickly reversed or entirely reestablished after collecting real-world data from participants or users. This is also seen in *Study I* as it mentions an eReader designed to be ideal called Acola, which ultimately performs similarly to the worst-rated eReader evaluated.

Both these studies outlined the importance of user participation in vetting a design concept, which is vital in the iterative process of producing successful, accessible, and intuitive websites.

References

Gedeon, T.D., Zhu, X.Y. and Dhall, A. (2015) "Felt depression is different to algorithmic depression: A user experiment using an image processing depression dataset," Computer Science Technical Report, CSTR-2015-13, Research School of Computer Science, Australian National University.

Gedeon, T.D. and Rampaul, U. (2015) "Popular eReaders," Computer Science Technical Report, CSTR-2015-14, Research School of Computer Science, Australian National University.