Using SUS for Current and Future AI

Richard Farry[1]

[1] QinetiQ

SUMMARY

The System Usability Scale (SUS) was assessed for its relevance and ease of use for assessing an AI capable of human-like interaction. Participants used SUS to assess Outlook, a contemporary consumer-grade AI interaction partners (smartphone digital assistants), and human teammates as a proxy 'system' for future human-like AI interaction partners. The results show that participants considered SUS to be relevant and easy to use for contemporary consumer-grade AI interaction partners, but not for human teammates. However, there was no meaningful difference in their ability to apply SUS between contemporary digital assistants, human teammates, and an email client. Thus, SUS can be used effectively for all of these kinds of systems.

KEYWORDS

SUS, Usability, AI, Human-Autonomy Teams

Introduction

Current and future Artificial Intelligence (AI) systems, particularly those intended to interact with humans as part of a human-autonomy team, will need to be assessed for their usability. It is not known whether current usability assessment methods are or will be suitable to assess such systems, particularly as (or if) AIs become more human-like in their interaction roles and competence.

This study set out to investigate the suitability of the System Usability Scale (Brooke 1996) to evaluate contemporary AI interaction partners, and future AI systems. A key strength of SUS, and why it was selected for this study, is that it can be used to assess a broad range of systems from any domain (Stanton *et al.* 2005), due to its use of general and high-level statements for participants to respond to. Additionally, SUS is easy to administer, can be used with small sample sizes with reliable results, and is valid; able to discriminate between usable and unusable systems (Brooke 2013).

Method

The hypothesis to be tested was whether participants found the SUS as relevant and easy to use for current and future AI systems that are intended to interact in a 'natural' way with humans, compared to using SUS for 'classic' desktop software using a Windows Icon Mouse Pointer (WIMP) interface. As future AI systems with human-like interaction were not available at the time of carrying out the study, human teammates were used as a proxy for such systems. The study participants completed a questionnaire that included three rounds of SUS. In order, they were for 'classic software' (Microsoft Outlook; the email client used within the participants' organisation), current and commonly available AI-based Digital Assistants (Alexa, Google Assistant, or Siri), and future highly-capable and human-interaction-like AI systems (using Human Teammates as a proxy for these future systems).

Following completion of each round of SUS the participants were asked to respond to the following statements on a 5-point Likert scale (from 'Strongly Disagree' to 'Strongly Agree'): "The [SUS] statements were relevant to the system", and; "I found it easy to rate the system". The participants were then invited to provide feedback about their experience of using SUS for each system.

Results

The participants were asked for each of the systems whether they considered the SUS statements to be relevant. The results are shown in Figure 1 below. There were thirty-one participants in total, but only eighteen of the participants responded to the questions about Digital Assistants.

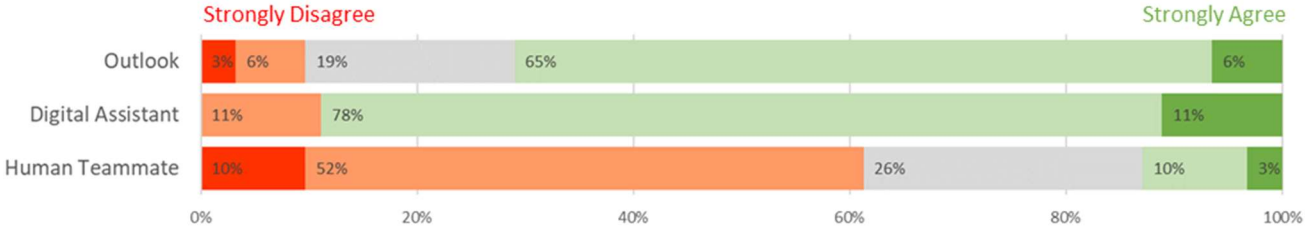

Figure 1: "Statements were relevant to the system" (n = 31, 18, 31)

The results indicate broad agreement that the SUS statements were relevant in the case of Outlook (71%, n = 31) and Digital Assistants (89%, n = 18). However, the participants considered the statements less relevant for their human teammates (agreement 13% and disagreement 62%, n = 31).

The participants were asked for each of the systems whether they found it easy to rate the SUS statements. The results are shown in Figure 2 below.

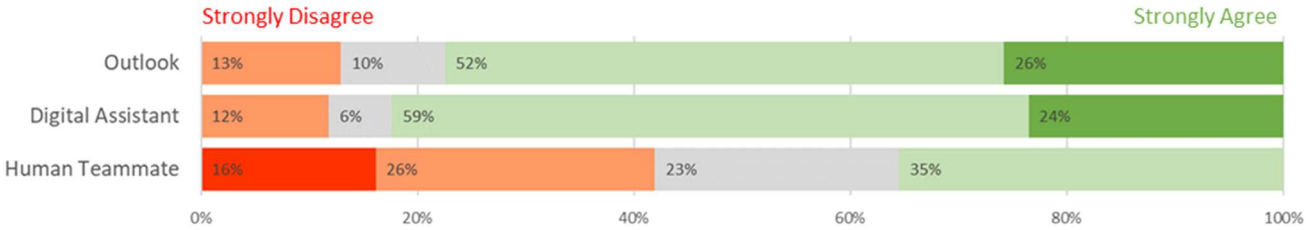

Figure 2: "I found it easy to rate the statements" (n = 31, 17, 31)

The results indicate broad agreement that it was easy to rate the SUS statements for Outlook (77%, n = 31) and Digital Assistants (82%, n = 18)1. However, there was less agreement and more disagreement on whether it was easy to rate the SUS statements when it came to their human teammates (agreement 35% and disagreement 42%, n = 31).

To investigate the objective usage of SUS a comparison of the proportions of 'Neither Agree or Disagree' ratings was carried out between Outlook and Digital Assistants, and between Outlook and Human Teammates. For this purpose an equivalence test (Lakens et al. 2018) was carried out, and in both cases the proportion of 'Neither Agree or Disagree' ratings were found to be equivalent

[1] The minor discrepancies between the percentages in the text and the figure for total agreement (agree plus strongly agree) are due to rounding.

(Digital Assistant: effect size tested = 0.1, Z = 2.483, $p < 0.01$. Human teammate: effect size tested = 0.1, Z = -2.147, $p < 0.05$).

SUS Scores

The overall SUS scores for each system are provided in Figure 3 below.

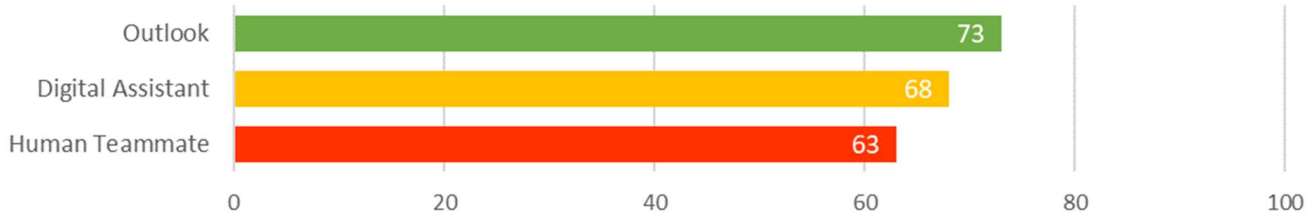

Figure 3: SUS Scores (68 is considered an average usability score (Lewis 2018))

Comments

The participants were asked to provide feedback on the use of SUS for each of the three systems. Twenty-four comments were received, and of these fifteen were about the use of SUS. The remaining comments related to the system being assessed. The comments about using SUS are summarised in the table below.

Table 1: Summary of comments about the use of SUS

System	Comment Type/Category
Outlook	SUS is too generic to capture relevant usability feedback (n = 3)
Outlook	Outlook has so much functionality [of varying usability] making it difficult to know how to respond to the questions (n =2)
Digital Assistant	It was more difficult to use SUS for a voice-based interface than a 'point and click' interface [i.e. a Windows Icons Mouse Pointer (WIMP) based interface] (n = 1)
Digital Assistant	The Digital Assistant is a front end to a range of functions / other systems, so it was unclear how to respond (n = 1)
Human Teammate	The SUS questions were difficult or not relevant to humans (n = 3)
Human Teammate	The SUS questions were impossible to answer about humans (n = 1)
Human Teammate	Neutral comment about the appropriateness of using SUS for a human (n = 1)
Human Teammate	Positive comment about the appropriateness of using SUS for a human (it was described as 'hilarious') (n = 1)
Human Teammate	Negative comment about the appropriateness of using SUS for a human (it was described as 'not appropriate' and 'demeaning') (n = 2)

Note that the participants who said rating the human teammate was difficult or impossible to do all successfully completed SUS, though their selection of the 'Neither Agree or Disagree' ratings (which could indicate difficulty in responding or simply just giving a neutral rating) accounted for 30.6% of their answers. The overall rate of participants responding with 'Neither Agree or Disagree' was 17.5% for Outlook, 15.3% for Digital Assistants, and 21.3% for Human Teammates.

Discussion

The participants rated Outlook as more usable than Digital Assistants and Human Teammates. This is perhaps not surprising in that Outlook is a tool designed to be usable, and is understandable and predictable in terms of its design and intended function (using what the philosopher Daniel Dennett refers to the as the 'design stance' (Dennett (2009)), whereas teammates are not, are far more complex, and unlike tools have other interests and goals.

Overall the participants considered SUS to be a relevant and easy to use tool to assess contemporary AI interaction partners (in the form of Digital Assistants), but not human teammates (as a proxy for future AI interaction partners). However, it was found that their ability to respond positively or negatively to the SUS statements for both was equivalent to using SUS for 'classic' software (in this case, Outlook). Thus, while subjectively they did not consider SUS to be valid, in practice their use of SUS demonstrated that it is an effective tool to assess AI systems, including those capable of human-like interaction.

Of more concern are some of the negative comments received about referring to people as systems, including one participant considering it to be 'demeaning' (see Table 1). It is unclear at this time whether similar concerns will arise for future AI systems, but it seems likely that they will if such systems are sufficiently anthropomorphic (or zoomorphic) or promote emotional engagement or attachment. This negative aspect of the use of SUS might be ameliorated with an appropriate briefing or introduction

SUS should be considered an appropriate means to measure the usability of contemporary AI interaction partners, and an effective stop-gap for measuring the usability of more advanced AI systems.

References

Brooke, J. (1996) 'SUS: A "quick and dirty" usability scale', in Jordan, P., Thomas, B. and Weerdmeester, B. (Eds.), *Usability Evaluation in Industry*, pp.189-194. Taylor & Francis.

Brooke, J. (2013) 'SUS: A Retrospective", *Journal of Usability Studies*, Vol. 8, No.2, pp.29-40.

Dennett, D. (2009) 'Intentional Systems Theory'. The Oxford Handbook of Mind, p339-350.Lakens, D., Scheel, A., and Isager, P. (2018) 'Equivalence Testing for Psychological Research: A Tutorial', *Advances in Methods and Practices in Psychological Science*, Vol.1, No.2, pp.259-269.

Lewis, J. (2018) 'The System Usability Scale: Past, Present, and Future', *International Journal of Human-Computer Interaction*, Vol.34, No.7, pp.577-590.

Stanton, N., Salmon, P., Walker, G., Baber, C. and Jenkins, D. (2005) *Human Factors Methods: A Practical Guide for Engineering and Design*. Ashgate.