Achieving Usability: Looking for Connections between User-Centred Design Practices and Resultant Usability Metrics in Agile Software Development

Adam Nemeth¹*, Anara Bekmukhambetova²

¹ Department of Ergonomics and Psychology, Faculty of Economic and Social Sciences, Budapest University of Technology and Economics, 2 Magyar tudósok körútja, H-1117 Budapest, Hungary
² Doctoral School of Business and Management, Corvinus University of Budapest, 8 Fővám tér, H-1093 Budapest, Hungary

* Corresponding author, e-mail: nemeth.adam@gtk.bme.hu

Received: 23 May 2022, Accepted: 15 August 2022, Published online: 21 December 2022

Abstract
This article searches for correlations between the practices of usability professionals from software development teams and the resulting usability of the system they are working on. After reviewing the literature of the past decade, we can conclude that most researchers try to achieve better usability with process proposals, but measurements focus on perceived usability. The level of involvement of users in the requirements capture phase is unclear. There is a trend to reduce costs of usability engineering, either by choosing more lightweight and informal methods or by trying to substitute user involvement. Looking at decade’s worth of 452 articles about “agile AND usability” in Scopus, we selected 133 which deal with the integration of usability, UX and User Centred Design practices and Agile development.

Keywords
usability, agile development, system usability scale, user-centred design, literature review

1 Introduction
Which factors within a software development team contribute most to better usability? Is better usability due to a certain practice, skill, or attribute of communication? Is it due to the team’s process? How ought we to measure the effectiveness of Agile and User Centred Development Integration (AUCDI) (Salah et al., 2014b)? This article will limit the concept of usability to the digital sphere, that is, the usability of information systems. We adopt the usability definition of ISO 9241-11:2018 as “effectiveness, efficiency and satisfaction” (ISO, 2018). Its importance lies either in the acceptance of technology (Venkatesh et al., 2003), if we treat “ease of use” as a synonym for usability, or the effectiveness of applying technology to a certain business process (Dishaw and Strong, 1999). As digitalisation becomes ever more prevalent, information system usability will also gain in importance, both on an economic and on a personal well-being level.

Numerous studies and frameworks exist on how to measure usability. It is widely accepted that the System Usability Scale (SUS) has proven to be a stable measurement of perceived usability that yields an absolute number (Lewis, 2018a). But perceived usability and actual usability might differ, especially when the user mistakenly believes that a task was executed correctly.

Where development practices are concerned, Agile (Beck et al., 2001) (as well as its Scrum implementation (Schwaber and Sutherland, 2017) became the leading software development process framework (Raunak and Binkley, 2017) in the mid-2010s. But as Sohaib and Khan (2010) puts it, Agile development does not necessarily result in usable software. Recently, continuous delivery practices have become more prevalent (Johnson and Mulder, 2020), but we will not include those in our discussions for now, as continuous operations, aka “Ops” behaviour (ResearchOps, DesignOps, etc.) is a relatively recent phenomenon, ResearchOps having their first discussions in 2018 (Towsey, 2018).

If Agile software development practices themselves do not necessarily improve usability, what does? Some (e.g. Gulliksen et al., 2003) argue that improved usability can be attributed to the adoption of User Centred Design (UCD) practices. The ISO 9241-210:2019 stan-
standard for such activities lists six principles of adoption (ISO, 2019a). But do they really improve usability? Others (Wale-Kolade and Nielsen, 2016) refer to maturity models such (e.g. Nielsen, 2006). Is there a measurable correlation between perceived and actual usability and the adoption of such practices?

We have looked through numerous tertiary studies (Curcio et al., 2019; Magües et al., 2016a) and after visiting Scopus, of the 452 English articles found relating to "agile AND usability" we have selected 133 for closer inspection between 2010–2020 to answer the following questions:

- Is there a clear correlation between integrating UCD into the Agile development process and usability?
- If the answer to the above question is yes, which UCD features are the most important, that is, which have the strongest effect on usability?
- How are usability of the end product, and adherence of the development process to UCD measured?
- Of all the proposed methods involving integration of Agile and UCD practices, which ones are measured objectively, and what do these measures tell us?

Which integration method leads to the best results?

2 Relevant literature

2.1 Scientific background

2.1.1 Relevant standards

We will base our definitions on parts of the ISO 9241-11:2018 and the ISO 9241-210:2019 standards, namely: Part 11, for definitions, and Part 210, for criteria of "Human-Centred Design" (HCD) (ISO, 2018; 2019a), a term which we will use interchangeably with UCD, as it was updated in 2010 based on the ideas of Gulliksen et al. (2003). The principles of part 210 concern objective understanding of user objectives and context, the inclusion of users during design and development phases, iteration based on user-based evaluation, and being able to look beyond the actual design artifact and the multiple viewpoints present in the design team.

In this present paper, we did not set standards criteria for Agile software development: anything which claimed to be "agile" in the literature was taken to be agile, if it was clearly about software development (as opposed to applying agile methods to construction or social care).

2.1.2 Costs and benefits of usability

Naturally, software development activities have associated costs in terms of worker time and equipment. The cost-benefit analysis is the topic of a fundamental book by Mayhem and Bias (Mayhew and Bias, 2005).

Based on the Task-Technology Fit model (Zigurs and Buckland, 1998), usability can be thought of as a measure of this fitness or of effort in terms of worker costs. Whether introducing technology is worth the effort can be thought as something which relates to the simplest version of the Technology Acceptance Model (Davis, 1989), where the perceived benefits of using new software might outweigh the difficulties associated with usage or vice versa.

2.1.3 Measuring usability

We will look at three aspects of usability: Satisfaction, Efficiency and Effectiveness, as well as how to measure each of them. We will detail literature on satisfaction here, and will discuss the objective measures in the discussion part, although a detailed analysis can be found in the works of Nigel Bevan (Bevan, 1995).

Satisfaction is and always will be that of a subjective metric, while the other two might be measured objectively. Satisfaction can be measured through perceived usability, by post-usage questionnaires such as System Usability Scale (SUS), Usability Metric for User Experience (UMUX) or UMUX-LITE (Borsci et al., 2015). A lot of research went into SUS (Lewis, 2018), and it is known to be stable and valid (Bangor et al., 2008). Moreover, it is stable across a set of languages (Gao et al., 2020), and according to Borsci et al.’s (2015) research the above metrics yield the same magnitude. Given the overwhelming scientific evidence, this paper assumes SUS to be a reliable way to measure the perceived usability of a system.

There are other ways to measure perceived usability, for example Questionnaire for Website Usability (Aziz and Kamaludin, 2016). Based on the low frequency of mentions in the reviewed articles we will ignore them for the rest of the discussion.

2.2 Literature review

2.2.1 Inclusion and exclusion criteria

We have searched Scopus for "agile AND usability" and selected articles between 2010-2020 for closer inspection. The date 2010 was chosen as the first publication date of ISO 9241:210:2019 (ISO, 2019a) as chosen as the most recent finished year.

Out of the 452 English results we removed 50 articles based on their abstract which did not deal with usability practices in software development but used the terms in an entirely different context (e.g. usability of an approach to a problem, or usability of agile methods in construction). We removed 5 articles which were not in English despite
the applied filter, and 67 conference proceeding headlines as relevant articles from them were listed separately as well.

We then removed 28 articles which used the term "usability" as a quality attribute without conducting any research to prove that it was improved, without making any effort to measure it according to any ISO-compliant definition, or without conducting any HCD-compliant activity to demonstrate improved usability. The process did not have to adhere to the ISO 9241-210:2019 standard fully (ISO, 2019a), but at least some efforts had to be made, such as involving users during design and development, making an objective assessment of the context of use, iterative product development based on user feedback, etc.

Next, we removed 167 articles which only described a single project or company case study (usually a software development project), and which did not deal with the cause-and-effect process whereby usability practice improved the usability results. We assume that although these efforts were made to improve the usability of the output, the focus was of these articles was not the integration of UCD and Agile.

Consequently, only 133 articles remained that genuinely focused on improving the usability of the delivered output in agile software development. This set of papers included tertiary studies, systematic mapping, and literature reviews.

2.2.2 On the articles left out: patterns in case studies
Many software development case studies used a subjective measure (either System Usability Scale or another survey) to measure perceived usability and made no effort to measure objective effectiveness and efficiency values. When measurement was done, instead of coming from usability tests or usage analytics instead it was based on expert evaluation, either using heuristics (such as Nielsen's (Nielsen, 1994)) or using a specific, sometimes locally invented (and probably highly subjective) method. This is likely due to lack of resources and lack of longitudinal data.

2.2.3 Measuring integration of usability practices
UCD practices are expected to maximise usability. How much they are integrated can be measured either by development team members filling out questionaries or by reviewing published practices either in case studies or proposed process models.

A proposed instrument is the Index of Integration (IoI) questionnaire by Joshi and Tripathi (2008). This already contains weightings.

The correlation of integration and results become evident in the study of the developers of IoI (Joshi et al., 2010). However, the authors of that paper use their own usability metric called Usability Goal Achievement metric (UGAM) instead of SUS. The issue with UGAM is that it is customised for each project before measurement and therefore cannot be accepted as a universal measure to compare different projects.

Maturity models could be used to measure integration as a replacement of Joshi's IoI method, as has been argued by Salah et al. (2014a), but in the reviewed articles only one such model had been used in this way (Salah et al., 2015). An obvious solution would be to use the ISO 9241-210:2019 compliance checklist to measure maturity, but in the articles selected we did not see that (ISO, 2019a).

Another study (Velmourougan et al., 2014) checked the practices in different software development models, and checked the number of usability defects in six different projects developed with different models in mind. The authors go on to propose their own development models.

A systematic review (García-Mireles et al., 2013) was done to see how different practices and usability correlate. However, the authors of that paper concluded that the relevance of the – at that time – current papers was low and further study was advised.

2.3 Current proposed process models
Of all documents visited by a review (Magües et al., 2016a) more than 47% were about processes, while in our study, out of 133 articles 24 deal with model proposals, mainly concerning the integration of usability activities into the development process.

Multiple models (Al Ghanmi and Jamail, 2020; Conforto and Amaral, 2016) propose a gated-model where usability practices in effect precede development, while other models suggest that evaluation should be made of the final artifact (Bergquist et al., 2020; Teka et al., 2017; Weichelt et al., 2020). Some papers note that the inclusion of users from the conception phase changed the direction taken by development (Herman et al., 2018).

2.3.1 Factors affecting usability adoption
Introducing usability practices entails costs, but whether that cost it is justifiable for the development company or its client was more recently discussed by Aydin and Beruvides (2014). Essentially, the costs and benefits should always be viewed as part of an ecosystem, which is discussed for example in Haile and Altmann (2016).
Out of the 133 articles, 24 deal with usability (user) testing, which seems to be the main way users are involved in the software development process. What is lacking here is more important than what is there: except for UCD books such as The UX Book (Harton and Pyla, 2018) and requirement engineering articles as reviewed by others (Magües et al., 2016b) there is no mention of actively involving users formulating the original requirements in abstracts, instead user requirements are considered as given or are based on assumptions/heuristics. This approach inevitably leads to conflicts in fixed-price projects (Kropp and Koischwitz, 2016) where usability issues come up only at user testing. Despite the Agile term "user stories", requirements are rarely based on information coming directly from users (Cajander et al., 2013), and instead, usability professionals are expected to provide the user's perspective.

Besides there being a lack of user involvement at the formulation of the project, there is a need to substitute, or extend the number of usability professionals by training developers to do their tasks multiple times, as has been proposed most recently in Øvad and Larsen (2016), using pair programming techniques (Seyam, 2015) or else conducting a cognitive walkthrough within the development team (Grigoreanu and Mohanna, 2013) at least as a preparation. An obvious substitution technique is to provide guidelines (de Oliveira Sousa and Valentim, 2019; Lee et al., 2010; Suleri et al., 2019) or heuristics (Choma et al., 2015). Personas (Caballero et al., 2013) can also be used as representations of user groups, but it is unclear whether the persona is based on rigorous context of use research or essentially made up.

3 Discussion
3.1 Measuring usability
For satisfaction and perceived usability, or the subjective aspect, SUS seems to be adequate.

Much of the criticism the usability construct has received concerns the justification given for reducing efficiency and/or effectiveness, but not the actual loss itself. We will not deal about "why" there is a certain friction: rather, we will measure it, which may serve to quench the anger found in most critiques of usability evaluations (Borsci et al., 2019).

Measuring efficiency is traditionally time-based. However, it is not an absolute measure per se but rather, it works in comparisons and depends on how far we take the substitution between given products or methods. For example, in the case of a new horse paddle, it is more straightforward to measure it in comparison to other horse saddles, not to a car or telecommute as a way of transportation.

Efficiency as an aspect of usability refers to the proportion of energy gained and effort spent by the users.

Effectiveness is measured by number of errors and success rate – with the former not necessarily coming from the latter, as some errors are recoverable (but recovery time will affect efficiency nonetheless). The success rate might be simplified to the term of conversion, meaning out of 100 users who started an interaction, how many of them felt they got what they want, but not necessarily: if a web shop honestly tells the customer the item they were looking for is out of stock, and therefore a purchase is not made, is it less effective than the one where the operator needs to refund the purchase for not showing this information?

Efficiency and effectiveness sometimes work against each other. For example, the risk of not allowing the user to enter values which would be illegal to submit. While its error rate would be zero, it would also not give any explanation why the requested operation is unavailable and might take more time to fill out than one which temporarily allows error states with a warning (that is, until the form is submitted). Therefore, for the sake of correctness, recoverable errors should be primarily accounted into efficiency and only success rates should be measured directly.

There exists a separation of success between how a user believes it has succeeded the operation and whether the operation was indeed successful – a famous example would be the Three Mile Island nuclear accident, documented in (Norman, 2002), where a valve was falsely believed to be closed due to how it was displayed in the user interface. Also, the previous example of "out of stock" might be considered success by a usability practitioner, even if the user does not interpret it as such (as a purchase did not happen). Therefore, judging the result as a success in terms of whether the user succeeded might need to be reviewed by a third party after use.

In conclusion, it can be shown, that while efficiency and effectiveness are linked (because of recoverable errors, since error recovery takes time), effectiveness affects efficiency, and if success states are clear, usability can be objectively measured by time and success rate.

Consequently, if it can be established that a user contacted the system to perform a specific task, a post-usage survey might adequately measure satisfaction, while time spent on the task and success rate can be used to objectively measure all other aspects of usability by log analysis. However, user testing is still needed to understand the factors behind numerical data, and false positives should be factored into measuring success rates.
3.2 The cost of usability and its ignorance

Usability is contextual: as a measure of task-technology fit, it depends on the task, or to be more precise, the context of use (Bevan and Macleod, 1994). In this way, usability equals the cost of this fit, and it can be compared to the benefits the execution of the task provides.

If the context of use is not established well, and only assumed, then in real world usage, the costs can go up higher. Whether it is recognised depends on how the usability study is executed: based on instructions, people can execute tasks which they would never do in real life, but in that case, only an imaginary persona might be satisfied.

Therefore, it can be argued, that while hiring usability professionals and involving users bear an obvious cost, that cost can be covered by the value created of the task-technology fit.

However, looking at the case studies it seems that at least some software development teams are executing client ideas instead of answering to user needs arising from context. Client satisfaction might be derived from the adherence to these ideas instead of solving a problem for users.

3.3 Measuring integration: a preliminary side study

A direct way of measurement for integration would be using the ISO 9241-210:2019 standard's compliance checklist by running an audit (ISO, 2019a), taking each item as a potential factor, and trying to assign weights to them through factor analysis, as these are the weights we are looking for.

While Joshi et al. (2010) show that there is a strong linear correlation between their integration metric IoI and UGAM on 61 projects, it should be noted that they only applied self-report evaluation by professionals who filled out both questionaries. In addition, no no-factor analysis was done on which IoI factors contributed the most.

In preparation for researching an ISO audit based survey, a preliminary study was conducted on the adherence to the six principles of ISO 9241-210:2019 (ISO, 2019a) with \( n = 28 \) HCD professionals present on a non-academic conference (Nemeth, 2020), about the project which took the most time of their week (since many of them work on multiple projects) (see Table 1).

The obvious outlier here is the result on principle of the inclusion of users: the highest frequency of the answer "1" (essentially never) was found here. This might hinder objective empirical full-scale measurement of usability, as inclusion of users is required for that.

3.4 Process models

Most process models reviewed put an emphasis on "shifted" or "parallel" sprints (iterations), in essence, ensuring that design activities precede development. However, many papers – while running internal iterations – only evaluated the last iteration quantitatively using surveys, instead of quantifying each iteration and thereby measuring progress. Some papers only evaluated the prototypes, but not the final product. Average usability (68 SUS score) is rather easy to reach, and a lot of papers did not achieve "A" scale on SUS (above 80), yet the iteration was stopped.

4 Summary

In conclusion, measuring usability essentially depends on three numbers: the SUS score, time spent to reach a pre-defined finish state of a task from a certain start state, and the ratio of users succeeding in it. However, it is highly dependent on context of use which should be established first, and critical errors might arise without the immediate understanding of the user. However, this is a client issue and sometimes not a concern for developers.

4.1 Answers to the original questions

There is an obvious correlation between integration of UCD practices as measured by IoI and its resulting usability, as measured by UGAM in multiple studies. However, these measures have their faults.

We could not find a factored analysis on which parts of the integration contributed the most in the Scopus literature of the past decade and this knowledge gap should be a basis of a further study.

Table 1

<table>
<thead>
<tr>
<th>ISO 9241-210:2019 principle</th>
<th>Average rating (scale = 1–5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Explicit and objective understanding of users, their tasks, and environment</td>
<td>3.1</td>
</tr>
<tr>
<td>2. Inclusion of users during design and development</td>
<td>2.1</td>
</tr>
<tr>
<td>3. Design process was iterative</td>
<td>3.7</td>
</tr>
<tr>
<td>4. Refinement was based on user-based evaluation</td>
<td>2.8</td>
</tr>
<tr>
<td>5. Design process considered factors behind the design artifact</td>
<td>2.4</td>
</tr>
<tr>
<td>6. The design team was multidisciplinary</td>
<td>2.9</td>
</tr>
</tbody>
</table>
Usability of the end-product is traditionally measured through SUS and usability tests. No longitudinal measures about conversion and time spent on tasks in actual, real-world, unattended scenarios were available. This can again, be a possible basis for further studies.

Most integration model proposals are either based on a case study (i.e. the author participates in or observes a development organisation which applies a certain integration, or a development project was prepared to study the effects of a proposed model), or based on surveying or interviewing usability practitioners and deriving results. To obtain an objective measurement, we need to measure not only the usability of direct competitors, but also the development practices of their respective organisations, which is a further study opportunity.

4.2 Further opportunities
Measuring the integration of usability practices could be done either as a secondary review of projects where SUS has been delivered, or surveying software development practitioners. This should be a topic of a further study, as results indicate as we could not find any study using ISO 9241-210:2019 (or ISO 9241-220:2019) as a base of checklist, and instead either resorted either to IoI or a maturity model (ISO, 2019a; 2019b).

While this seems straightforward, no factored analysis concerning which usability practice contributes the most was found.

The absence of users during the design process is a warning, as HCD (IxD) professionals might not be able to establish the context of use correctly, and as such might measure usability in relation to the wrong tasks, and might not get adequate, real-world feedback on their work in terms of usability. Moreover, only the customers of software development decide whether their product is acceptable, without ever seeing its real-world usage. How this affects usability could be a further study in itself.

Also there seems to be a problem on a divide between usability professionals and developers (either the prototype is measured or the end product, rarely both), and for some reason, quantification of user experience between iterations of the product / prototype using the same survey is also rare. Exploring the reasons of this might also need a separate study.

Acknowledgement
The authors would like to thank for Karoly Hercegfi, Andras Nemeslaki, and especially Andras Nagy for adding suggestions to the preliminary documentations of this study.

References


