Tampere University

Saila Juuti

HEURISTICS EVALUATION OF LEARNING ENGAGEMENT BASED ON PSYCHOLOGICAL NEED FULFILLMENT MODEL

Faculty of Information Technology and Communication Sciences

Master's Thesis

April 2020

ABSTRACT

Saila Juuti: Heuristics evaluation of learning engagement based on psychological need fulfillment model
Master's Thesis
Tampere University
Master's Degree Programme in Internet and Game Studies
April 2020

This thesis both documented the process of creating the Learning Engagement Evaluation method for evaluating digital and physical learning solutions and presents a research for its additional validation. The creation of literature review on the topic describes the theoretical foundation of the method, and the thematic analysis of problems the teachers encountered while conducting the evaluations with the method provided further validation and insight to points of improvements.

The developed Learning Engagement Evaluation is a heuristic expert evaluation method using six basic needs suitable for UX design: autonomy, competence, relatedness, respect, stimulation and safety. These were adapted from the work of Hassenzahl and colleagues (2010). Need fulfilment has been applied to analysis of user experiences both in UX research and in games research and heuristic frameworks based on need fulfillment have been developed previously.

To further validate the research, 42 evaluations conducted by teachers to 14 learning solutions were analyzed to find misunderstandings and problems when applying the heuristics. The conclusions of the study are that some concepts need more explaining and preferable examples of applying them. In general, the method was found solid and well applicable.

Keywords: heuristic evaluation, learning engagement, self-determination theory, educational technology, user experience, game studies

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

TIIVISTELMÄ

Saila Juuti: Heuristics evaluation of learning engagement based on psychological need fulfillment model
Pro gradu -tutkielma
Tampereen yliopisto
Master's Degree Programme in Internet and Game Studies
Huhtikuu 2020

Tässä opinnäytetyössä dokumentoidaan prosessi, jossa luotiin menetelmä oppimisen sitouttamisen (Learning Engagement) arviointiin, sekä esitetään tutkimus sen lisävalidoinniksi. Aiheesta tehty kirjallisuuskatsaus kuvaa menetelmän teoreettista perustaa, jonka lisäksi tehtiin teema-analyysi ongelmista, joita opettajat kohtasivat käyttäessään menetelmää oppimisratkaisujen arvioinnissa.

Kehitetty Learning Engagement -menetelmä on heuristinen asiantuntija-arviointimenetelmä digitaalisten ja fyysisten oppimisratkaisujen arviointia varten. Menetelmässä käytetään kuutta UX-suunnitteluun sopivaa psykologista perustarvetta: autonomia (autonomy), kompetenssi (competence), yhteenkuuluvuus (relatedness), käyttäjän kunnioitus (respect), stimulaatiota (stimulation) ja turvallisuutta (safety), perustuen Marc Hassenzahlin ja kollegoiden (2010) työhön. Kirjallisuuskatsauksessa todetaan, että psykologisten tarpeiden tyydyttämistä on laajasti sovellettu käyttökokemuksen analysointiin sekä UX- että pelitutkimuksessa, ja siihen perustuen on kehitetty myös muita heuristisia menetelmiä.

Menetelmän validoimiseksi analysoitiin 42 arviointia, jotka opettajat suorittivat 14 oppimisratkaisulle. Tutkimuksessa analysoitiin väärinkäsityksiä ja ongelmia heuristiikkojen soveltamisessa. Tutkimuksen johtopäätökset ovat, että jotkin käsitteet tarvitsevat enemmän selitystä sekä esimerkkejä niiden soveltamisesta. Yleisesti menetelmä todettiin toimivaksi ja helposti sovellettavaksi.

Asiasanat: heuristinen arviointi, oppimisteknologia

Table of Contents

1 INTRODUCTION

"Let me introduce the word "hypertext" to mean a body of written or pictorial material interconnected in such a complex way that it could not conveniently be presented or represented on paper. It may contain summaries, or maps of its contents and their interrelations; it may contain annotations, additions and footnotes from scholars who have examined it. Let me suggest that such an object and system, properly designed and administered, could have great potential for education, increasing the student's range of choices, his sense of freedom, his motivation, and his intellectual grasp. Such a system could grow indefinitely, gradually including more and more of the world's written knowledge."

Theodor Nelson (1965)

In the paragraph above Nelson, a pioneer in information technology and the inventor of word "hypertext", is describing the potential of interactive technology for creating new kinds of learning experiences. Half a century later, we can see that he was correct in many ways; technology has become ubiquitous, and all information is linked, indexed and within our reach. But has learning changed? Are the students more self-directed, motivated and intellectually-intrigued, as Nelson would predict?

We know that digital tools for learning are available, and they are already widely-used. A study by the European Commission (2nd Survey of Schools: ICT in Education, 2019) states that in the EU, 68% of students at ISCED Level 2 and 73% of students at ISCED Level 3 use the Internet at least once a week for learning purposes, either with their own devices (laptops, computers and smartphones) or by devices owned by their school. Based on the same report, students are using digital tools for producing content (editing text, images and multimedia), accessing digital resources such as books, simulations and games, and doing learning exercises and tests.

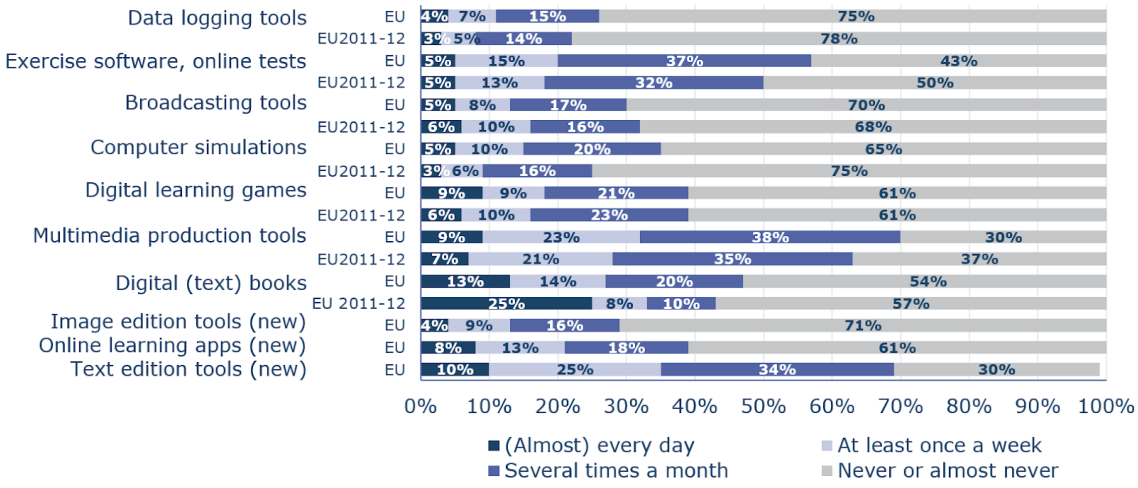


Figure 1 The use of digital resources and tools during lessons
(ISCED 2, in % of students, EU level, 2011-12 and 2017-18), European Commission, 2019a.

European Commission also sees digitalization of education worth promoting, and has set an action plan to support technology use and the development of digital competences in education, the focus of which is to improve the infrastructure in schools and the skills of the users (Digital Education Action Plan. Accesses 8.3.2020).

The European Commission also sees the digitalization of education as worth promoting, and has set an action plan to support the use of technology and the development of digital competences in education, the focus of which is to improve the infrastructure in schools and the skills of the users.

However, not as much focus has been put on to materials and resources. Several market studies state that the market for digital educational solutions is fragmented; there are large technology companies, such as Google, Apple, and Microsoft offering their ecosystem for education; traditional educational book publishers are creating digital materials; new emerging companies are offering individual learning solutions on various topics. Nevertheless, it is not easy to assess the quality and impact of each of these solutions, nor the learning experience they offer to the student.

To answer this need, Education Alliance Finland (EAF), a Finnish expert organization, has created a method for evaluating the pedagogical quality of different kinds of digital and physical learning solutions. EAF is a privately owned company, that employs professionals on pedagogy and interaction design for conducting the quality evaluations. All evaluations are done as expert evaluations, where the evaluators use the solution under review and assess it using the EAF method. The evaluations have two main parts: a pedagogical evaluation and a learning engagement evaluation. The pedagogical evaluation identifies the learning goals the product supports and assesses the pedagogical value by evaluating what the student's role is when engaging with the product. The theoretical framework of the evaluation is based on the Engaging Learning model (Lonka and Ketonen 2012; Lonka 2012), and is described in a white paper article by Hietajärvi and Maksniemi (2017).

The topic of this thesis considers the second part, the Learning Engagement Evaluation. It aims to holistically assess, if the learning product gives the user a positive experience which would encourage the further use of the product and provide a good learning experience. Despite the name, the Learning Engagement Evaluation does not assess if any learning can happen during the use; even if high pedagogical value and positive user experience often overlap, it is possible for a product to gain good results in Learning Engagement Evaluation even if the pedagogical value is poor. However, a product with poor user experience is very likely to have poor educational value as well.

When conducting this research, as a co-founder and employee of EAF I am in a dual role; while working in the company, I was responsible for creating the Learning Engagement Evaluation method. Work started in 2016, and the process continued iteratively up to the year 2019. When the method had found its current form and it had been validated in several ways, EAF developed software that helped to streamline the evaluation work and automatize parts of report creation. I was the designer of the software and while consulting the rest of the team made decisions on the included features and page layouts. In this thesis I am both documenting the method creation process and further validating the framework, as well as discussing the flaws and developmental points both in the method and in the software. In the last part of this study I will analyze how teachers have used the method and if further iterations would be required.

I position my research in the field of Human-Computer Interaction (HCI) research and games research. HCI is a field of study focusing on different aspects related to the use of digital solutions. User experience (UX) and usability research and heuristic evaluation methods have their roots in HCI research, although it is possible to assess UX and usability of non-digital solutions as well. The learning engagement framework is based on Marc Hassenzahls work on psychological need fulfilment as a basis for user experience evaluation and design (Hassenzahl 2010, Hassenzahl et al. 2010, 2013).

Although the topic is the assessment of learning solutions, this research is not in the field of pedagogy, as the learning engagement assessment will not take a stand on how much or how effectively the user is learning - only what kind of emotions the learning experience with the evaluated solution will evoke. For the literature review I also will consult academic sources on educational psychology, as there are many ways that how research on game experience overlaps with learning theory. The nature of game research can still been seen as interdisciplinary (Crookar 2000; Deterding 2017), and the researchers in the field come from various backgrounds (Mäyrä, Looy, and Quandt 2013).

2 THE MAIN RESEARCH QUESTIONS

This research tests and aims to validate the heuristic framework for assessing the user experience of a learning solution, which is used in EAF evaluations. The framework is used by teachers, who are experts in pedagogy and classroom practices. However, they are not experts in user experience evaluation, although they are trained to use the methodology. The framework should be flexible enough to suit a variety of products and analyzing products with it should make sure that all relevant aspects of the products have been notified and a holistic view has been formed. In particular, this study addresses the following questions:

- How to use the need fulfilment as a theory for heuristic evaluation of user experience and learning engagement?
- How the teachers creating Learning Engagement Evaluations are using the heuristics and what are the possible points of improvements.

Although the framework is used for evaluating learning solutions, the heuristics are not focused on the aspects of learning or pedagogical validity, but instead, what properties of the learning solution give the user positive experiences that would encourage the further use of the product. It is possible to assess also products made for entertainment or free play with the set of heuristics, and it would be possible that a product with poor educational quality or very limited educational content would still align well with the learning engagement heuristics. It would be an interesting question to see if there's correlation between pedagogical validity and good user experience, but assessing that is out of the scope of this study.

2.1. Method and research data

A framework for assessing the user experience of educational solutions is already formed and used in EAF evaluations under the title Learning Engagement Evaluation. More than 200 products have been evaluated with the learning engagement framework between years 2016-2020, and the method has been iteratively developed based on its

applicability to different types of products. It is based on an existing list of heuristics from the fields of usability- and game research, which are modified to fit the context of analysing learning solutions. The heuristics are categorized according to the set of needs suitable for Experience Design as proposed by Marc Hassenzahl et al. (2010, 2013).

The aim of this study is to document the process of iteratively creating and applying the heuristics, and further analyze its validity and future development needs by comparing the process, the framework, and the experiences of using it for the existing research. The study includes a literary review of user experience assessment with expert evaluation methods (not including methods based on user testing or user observation). The literature review discusses the specifics of assessing learning solutions, but will not be solely limited to those. The review will use theoretical background from the fields of HCI and games research, as well as psychology for the related parts.

The research data used in the second part of the thesis is 42 evaluations conducted with the method. In each evaluation, the evaluators give a rating for every heuristic in the framework based on how they see the product. They can also ignore the heuristic if they think it is not relevant for the current product. The evaluators can also leave written feedback. The selected evaluations were analyzed with qualitative methods to understand how useful the framework is for assessing the user experience, and how it could be improved. The evaluations were all conducted from January to April 2020 by mainly novice users of the method. Despite the great number of evaluations conducted before 2020, the evaluations done under the data collection period provided documented evidence on potential problems. Most of these problems had been previously recognized intuitively and addressed case-specifically, but they were not formally documented or analyzed prior this study.

3 LITERATURE REVIEW

In EAF evaluation the heuristic evaluation of user experience is called Learning Engagement Evaluation. The term was selected to describe the subject and goals of the assessment; It is conducted on products made for educational purposes and the method itself includes an assumption that a positive (ie. meaningful), user experience leads to better learning engagement.

The method is used by experts to define, if the experience the product delivers will be engaging for the actual users of the products - may they be children, higher education students, casual adult learners or learners in corporal contexts. The fundamental question this approach raises is whether it is possible to make statements about other users' experience by examining the product itself, and especially, if it is possible to say anything about the learning experience of the user. The topic has a philosophical and epistemological component to it, and Human Computer Interaction (HCI) research has examined it from various practical points of view.

3.1. What is engagement?

Engagement can be approached from various academic perspectives. When discussing the use of products and services, engagement is often understood as user engagement, and has been studied also by researches on media, marketing and advertising (Oh, Bellur and Sundar 2018) For this study, HCI research produces the most relevant viewpoint. However, understanding of engagement also in HCI is based on the psychological theories of motivation (Hassenzahl 2010, 11, 40). Many of them are also applied in educational research when discussing student's engagement with school or learning activities. In the field of educational research, engagement, (also called *learning engagement* or *student engagement)*, is a concept that has various meanings

that can be sometimes seen as vague and confusing to distinguish (Ashwin and, McVitty 2015), so I will shortly present how the concept is used in that context and present a definition of learning engagement as it is used in this study.

3.1.1. Learning engagement in educational research

Engagement in the context of learning has been studied by educational psychologists and pedagogist, and it is recognized as a factor for academic success and increased learning outcomes (Kuh et al. 2008; Svanum and Bigatti 2009) However, the terminology around engagement is not completely unified, and the pedagogical research uses the terms school engagement student engagement and learning engagement often interchangeably (Appleton et al. 2008). In the EAF method and this study the term "learning engagement" was selected because it also covers learning solutions not used at schools nor by "students" - rather, the only thing implied is an existence of a *learner*. In this chapter I am shortly presenting one taxonomy of understanding engagement. Since the EAF Learning Engagement framework doesn't claim to make statements about learning, I will present this view only to show how it is related to theories of motivation and that way to principles of product design that the EAF framework is following.

Fredricks et al (2004) consider engagements as a meta-construct and identify three different dimensions of it: behavioral engagement, emotional engagement and cognitive engagement. *Behavioral engagement* is related to a student's actual participation in the learning activities, *emotional engagement* to their feelings about learning and *cognitive engagement* to mental effort required for learning, setting of learning objectives and self-regulating learning activities.

The conceptual model of behavioral, emotional and cognitive engagement has been applied in and discussed in numerous studies. Prior to research of Fredricks, a two-point model, behavioral and emotional or affective engagement without addition of cognitive engagement has been suggested, and some models add a fourth component, academic engagement, that is separated from behavioral engagement (Appleton et al., 2008, Christenson et al. 2008). The three different modes of engagement have also been found to be interrelated(Li and Lerner 2012). Although personality traits of individuals are affecting engagement in educational settings (Shernoff 2013), engagement is also

seen as malleable and greatly depending on the context (Fredricks et al. 2004, Appleton et al., 2008), which suggest that external factors, such as study settings affect the student's engagement and that correctly designed learning activities would increase the learning engagement.

Following the three-point model of engagement, Fredricks et al. have also compared ways of measuring the students' learning engagement. In their meta-analysis they identify five methods; self-report measures, experience sampling techniques, teacher ratings, interviews, and observations (2012). What is important to point out is that all these methods are based on studying the students (not for example learning products), and often focus on overall engagement to academic activities, often in a formal learning setting.

For this study, the three-point model was used as the most relevant construct for understanding learning engagement with designed learning products, with focus on behavioral engagement and emotional engagement.

When interacting with a digital product for learning purposes, behavioral engagement is the most straightforward to observe - is the student interacting with the system or not. In this context, separating components of academic engagement from the other behavioral engagement could be irrelevant. According to Appleton et all (2008), when studying learning engagement in school context, academic engagement is measured with time on task, credits earned toward graduation, and homework completion rates, whereas behavioral engagement is related to attendance, being present in class or extra-curricular activities. With digital products, "attendance" could be seen as times the student has logged in and separating that from eg. task completion rates could provide interesting viewpoints and valuable information, but this couldn't necessarily be applied to all learning products without problems. For example, in solutions where the student explores the system freely and defines their own goals of use, or systems that are very focused on casual and informal learning, the separation feels very school centered.

The EAF whitepaper Designing Engaging Learning Solutions (Hietajärvi and Maksniemi 2017) discusses behavioral engagement and defines it through the mandatory interaction between the solution and the users: Does the solution require

active engagement (doing things) to progress or does the solution allow user to pass through the content with no/low engagement. A solution that requires the user to make inputs, for example click on things or solve problems to proceed, makes the user engage with it, at least on a behavioral level. However, this doesn't guarantee that the user is engaged on an emotional or cognitive level.

In several studies, cognitive engagement has been found generally difficult to measure or identify. It is most often related to self-regulatory activities such as setting of one's own learning goals and how much the students express valuing school work (Appleton et al. 2008).

To promote emotional engagement, a learning solution should make the learner interested in the usage or topic of learning and motivate them to solve problems and progress. According to Frederics et al. (2004) the research on emotional engagement "refers to students' affective reactions in the classroom, including interest, boredom, happiness, sadness, and anxiety". When engaged, the learner has interest towards the learning activities, and the interest can be seen either as personal or situated. A situated interest can be "aroused by specific features of an activity, such as novelty", whereas the personal interest is more stable (ibid).

In learning product design and when taking new learning technology into use, novelty factors play a role in students' engagement and interest to use, and therefore can increase the learning outcomes. This was tested by Jeno et al. (2019) for mobile learning and by Burke and James (2008) for use of PowerPoint as a learning tool. However, as was pointed out in both of these studies, the novelty vanes, and the benefits for engagement will decrease (see also Keller and Suzuki 2004).

The other feelings in the learning situation, such as boredom, enthusiasm, joy or wonder are naturally highly subjective and personal to an individual student. However, there are ways how these could be predicted by analyzing factors in the situation. I will discuss this further in the chapter 3.3 in the context of analyzing the user experience of learning solutions. As I will show, the student's potential engagement when using a learning solution can be (within limitations) assessed by analyzing the properties of the solution.

3.1.2. Definition of engagement in this study

The EAF learning engagement framework is targeted at analyzing learning engagement when using a certain learning solution and performing tasks within the limits of the solution. However, the learning solutions that can be evaluated with the framework are not necessarily to be used in formal educational settings, so definitions of engagement that are tied to schools or the presence of students and teachers as users are not fitting.

In their research on learning engagement in technology-mediated learning, Hu and Hu (2012) present a following definition for learning engagement: *"a student's voluntary participation in activities designed as part of the learning program*"

In this study, learning engagement is defined in this manner, although "student" could be replaced with "learner" or even "user". A robust definition brings up important points about engagement when using different kinds of designed learning solutions.

"Voluntary participation" means the student is willing to take part in the (designed) learning activities. The use of a certain solution can be "forced" in situations such as schools, but voluntary participation implies that the student is actively using the said solution for reasons other than being forced to do it. "Participation" brings up the point about interactivity. The EAF evaluation method is developed for solutions, where the learner is interacting either with software, physical objects or other people, and performing activities with them. It can be said that reading a book or watching a video as a part of an educational setting is participation. Yet, it is not meaningful to evaluate a book with heuristics developed for assessing interactivity.

By this definition, the aim of the Learning Engagement Evaluation is to *understand how different factors in the learning solution affect the learner's willingness to interact with the solution.* It uses the methods of HCI research to assess the user experience through analyzing the properties of the product. Also following the definition, the EAF learning engagement assessment will not take a stand on how much or how effectively the user is learning - only *what kind of emotions the learning experience with the evaluated solution will evoke;* mainly, if the learning product gives the user a positive experience which would encourage the further use of the product.

Following the three-point model of engagement, this suggests that Learning Engagement Evaluation is most interested in the emotional engagement of the students. Some factors that facilitate cognitive engagement can be identified with the method, for example relating to the types of challenges the products are giving to the user, but as the research on the topic concludes, this is difficult to measure without observing or interviewing the users (Appleton et al. 2008). The behavioral level of engagement is implied only with "participation" - in assessment of interactive products, participation can be expected because the product needs interaction to progress. However, measuring or assessing what the user actually does in the product and how much time and effort they invest in the use is not in the scope of the Learning Engagement Evaluation method, because the method doesn't involve following the behavior of real users. This would be an interesting metric in product development, and could be measured for example with click through rates of prompts or frequency of logging in.

3.1.3. Need fulfilment as a basis for engagement

Engagement is directly related to motivation (Pintrich 2003). In psychology, there are several theories and models on motivation and where it stems from. The Learning Engagement Evaluation model is based on the assumption that there are basic psychological needs, and their fulfillment can be supported by appropriate design in learning products. This need fulfillment in turn will then increase the user's motivation to the use - and if the learning solution is also pedagogically properly designed - to learning. The psychological needs used in the Learning Engagement Evaluation are based on the need fulfillment framework for UX design by Hassenzahl et al. (2013), who adapted it from the research of Sheldon et al (2001). This framework is strongly based on a model of motivation presented in self-determination theory (STD) by Deci and Ryan (2000a).

STD starts from an assumption that human beings have natural, internal tendencies to develop elaborate and unified sense of self, as well as "forge interconnections among aspects of their own psyche as well as with other individuals and groups in their social worlds" (Ryan and Deci 2002, 5). This need for self-development then manifestates in three basic psychological needs; need for *autonomy, competence, and relatedness.* The

need for autonomy reflects a desire to be in control of and self-determining in one's behavior, the need for competence refers to the desire to master and be competent in one's interactions, and the need for relatedness reflects a wanting to belong or be attached to a group (Pintrich 2003).

STD is focused on understanding a person's actions in relation to their social environments, and sees that certain kinds of environments can either support the fulfilment of these needs or hinder it. The needs are seen as universal and innate for all humans in all cultures, although STD also states that these needs can be fulfilled in variety of ways (Ryan and Desi 2002, 9)

STD has been widely used to analyze factors behind student engagement in educational settings (eg.: Jeno et al. 2019, Reese 2012), and the basic principles behind it have been found to be applicable. Or, as Reese (2012) puts it: "SDT is a theory of motivation that helps researchers and practitioners alike to understand and enhance not only student motivation but also the engagement that arises out of that motivation."

STD has been also applied to understand principles in product design that make the use of certain products motivating and interesting, as well as a guiding theory in a practical digital learning solution design (Ford, Wyeth and Johnson 2012). Besides being a foundational theory in the need fulfillment model of Hassenzahl at et (2006), which I am using as a background theory for Learning Engagement evaluation, it has been used in studying games and gamification. In a review of gamification research by Bozkurt et al. (2018, 27), SDT was the most common theoretical or conceptual framework that was used to analyze gamification.

In their study on the motivational effect of video games Przybylski, Ryan and Rigby (2010) noted that games have increasingly incorporated the aspects of SDT in their design. The competence element of challenges the game provides have been complemented with autonomy (such as free exploration of game worlds) and relatedness (support of online communities).

STD has been also used to suggest design guidelines for gamification and motivating user experience. According to meta-analysis by Mora et al (2017, 537), SDT was the

most popular motivational theory utilized by the gamification design frameworks. For example research by Weiser et al. 2015 linked design components found in persuasive technology (games and gamified systems), such as feedback and challenges, to theories on motivation, including STD. They presented a t*axonomy of motivational affordances*, and using a case study (project GoEco!), demonstrated how the taxonomy can be used as a design guideline for a persuasive gamified system. Also Proulx et al (2017) suggest a set of game design principles for educational mobile games that should feed to the player's intrinsic or extrinsic motivation according to STD.

Need fulfillment has also been used as a theory for several heuristic frameworks. I will describe those more closely in relation to my own heuristics in Learning Engagement evaluation in chapter 4. Next, I will discuss the ways how experiences and engagement in the use of technology can be assessed, and how motivational theories, specifically STD, can aid in that.

3.2. Technology mediated experiences

As presented in the last chapter, engagement can be studied in various contexts. In my study, I am interested in engagement while using different designed learning solutions, and if and how those solutions can foster and support engagement. In the context of this study, *learning solution* refers to designed, interactive products that are largely technological in nature, and are used for learning purposes. It would be possible to refer to these products as learning technology or educational technology (or Edtech), but "solution" covers also non-digital interactive products such as board games. However, since a great majority of modern interactive learning products use technology (software and hardware), it is best to understand their usage through the research concepts related to (educational) technology.

User experience research is a field of study interested in questions about reasons for using certain products, services and software, and what kind of experiences these give to the users. Although it currently covers a broad scope of research, user experience research has its origin in the field of Human Computer Interaction (HCI) research, which is an interdisciplinary field of study focusing on different aspects related to the use of digital software and technology (Grudin 2012, xxviii). HCI research has traditionally focused on goal fulfillment: what the user is trying to achieve when using

the technological solution (Hassenzahl 2010, 13). When understood very narrowly, the goal is seen as something the solution is designed for - for example, a person wants to send a text message by using their phone. When understood in this manner, the experience is assessed based on the features in the technology, and how they serve the goal (ibid). This is a valid viewpoint for example when assessing the usability of the product.

When using technology, people are experiencing a variety of things, which might or might not all be related to the features of the technology. In HCI, the concept of *User Experience (UX)* tries to capture this. Jesse James Garret defines user experience as "the experience the product creates for the people who use it in the real world" (Garret 2003, 10). Or, as Hassenzahl (2010, 8) puts it, user experience focuses on how "interactive products create, facilitate and mediate experiences for the users".

The UX point of view acknowledges that technology is always used in a context; the personal knowledge, abilities, feelings and motivations affect the way it is used by different users and how the usage is experienced. In addition to that, the experience is always somewhat subjective and unique to a given situation. Although a designer can try to design for a certain kind of experience, the experience can't be guaranteed. Rather, the real experience emerges from factors that are only partially under the control of the designer. (Hassenzahl 2010, 6, 10-11)

In this study, the "end user" refers to the learner, whether they are students in an institution or a learner studying in their free time at home. The end user can also be the teacher using the product with their student. It should be recognized that the learner and teachers may have a very different user experience with the same product. In the context of this study, it is useful to think of the user experience as a learner's learning experience that is mediated by learning technology.

3.2.1. Technology mediated learning

Like Theodor Nelson in his prediction about the potential of hypertext in education (1965), many scholars and educators have seen technology as a valuable tool for

learning. Alavi and Leidner (2001) present definition for *technology-mediated learning* that "refers to an environment in which the learner's interactions with learning materials, peers, and/or instructors are mediated through information technologies". This covers various ways how technology can be used in learning; use of technological devices and software for various purposes.

In educational research a central question related to the use of technology is its impact on learning outcomes. In early studies of learning technology, some researchers presented ideas that adoption of new technologies and emergence of new media formats, for example use of video, has attributes that in themselves makes learning different and more efficient. In his response to this conversation, Richard E. Clark (1994) makes a claim that new media technologies don't bring any "unique contribution" to learning.

Clark's argument is that any media is simply a vehicle of learning, and "it cannot be argued that any given medium or attribute needs to be present in order for learning to occur". He sees that studies showing the use of new media having a positive impact of learning should be put to a *replaceability test* to see if similar learning outcomes could be achieved with a setting that is not using the said media. (Clark 1994)

Clark doesn't claim that technology wouldn't have any benefits at all; use of technology can make learning more cost effective and accessible, certain media or attributes can be more efficient for some learners. There are also things which can be best, or perhaps only, be learned by certain media. For example, learning programming or video editing can be taught without technology, on pen and paper - but this would be slow and inefficient. Topics that are essentially about the medium are best taught with the medium. Clark's critique is towards lifting media above the message - any technology that is used should be correctly designed to deliver pedagogically valid content. Therefore Clark's view can be seen as a rather clear-headed response to the hype of bringing new technologies in learning at that time.

In 1994, Clark couldn't predict how ubiquitous technology would become; use of communication technologies for work and leisure has become an integrated and potentially irreplaceable part of our practical daily activities. Yet, similar myths about technology fostering motivation and learning outcomes are still present in the discussion

(Kleiman 2000), and as Whitton (2010, 596) states, even researchers may view educational games having motivational effects to learning as such, regardless of their content.

Since then, many studies show that correctly designed use of technology has potential for creating experiences that can be motivating and represent very solid pedagogy, and that way improve the learning outcomes (for a meta-analysis, see Chauhan 2017) . As presented in the previous chapter, experiences with technology are always contextual. In educational research, the context of use has been studied and conceptualized especially in relation to using technology in schools. Wan, Fang and Neufeld (2007) present a model of technology mediated learning that sees technology in relation to student, teacher and the instructional design setting; technology can be used to deliver content or as a method of interaction between teachers, students and peers. The model of Wan, Fang and Neufeld posits that information technology affects students' learning outcomes in interaction with the people that are involved and the overall design of the learning setting. This is in line with other studies that see for example the teacher's role and attitude as crucial in the successful use of learning technology (Piccoli et al 2001, also European Commission 2019a, 2019b).

A further conceptualization of the context of usage of learning technology is presented in "The 2nd Survey of Schools: ICT in Education" - report by the European Commission (2019b). A model of "highly equipped and connected classroom" is based on the study findings of extensive research on the use of technology in schools in European Union (ibid.).

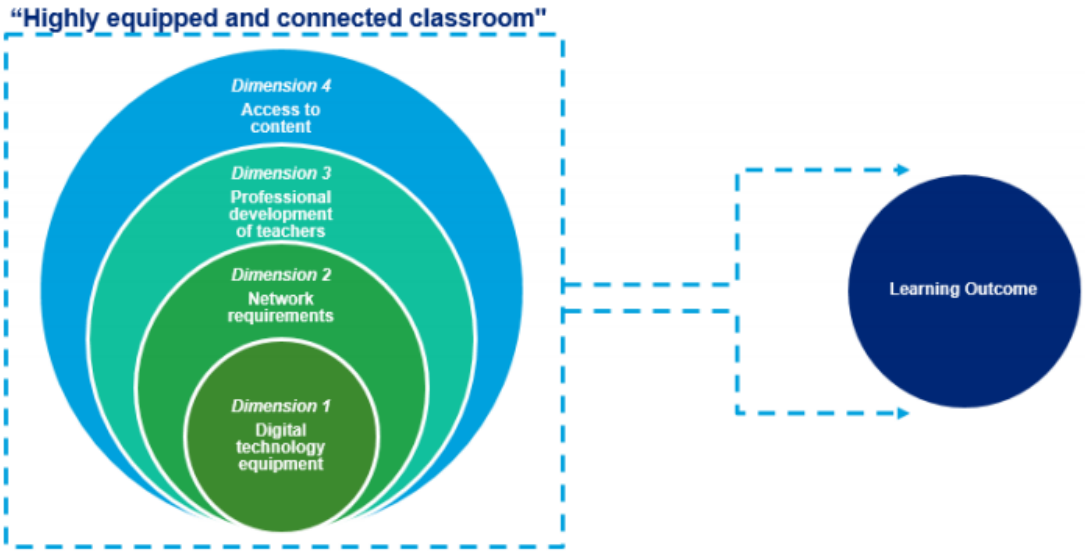


Figure 2: 'Highly equipped and connected classroom', European Commission (2019b)

The model suggests that in order to use technology efficiently in education, the infrastructure (equipment and network) need to be sufficient, and the teachers need to have a required skill level in the usage. The quality and use of content mediated by educational technology should be considered only when these factors are in order. The concept of "educational technology" is defined broadly: "technologies that are used in educational settings for learning and teaching purposes" (European Commission 2019b). This includes both physical technologies (i.e. hardware) and educational software and services, including learning management systems, software tools such as word processors, and software delivering educational content, like e-books or learning games.

These kinds of frameworks help understand the context of use of educational technology. In EAF Learning Engagement framework the evaluators are teachers, so they should have practical understanding of social and infrastructural realities in schools, and therefore are able to utilize it in their evaluations.

3.3. Need fulfillment in analysis of user experience

I have presented that the (learning) experience with technology is always situational, contextual and subjective (Hassenzahl 2010, 6, 10-11), and therefore a designer can only partially control, what kind of experience a product is mediating. Yet, Hassenzahl et al concludes that ”The assumption that need fulfilment leads to a positive experience is apparent in a number of models of user experience in the context of HCI.” (2010), and point out that and that designers should consider what need an interactive systems is fulfilling, not just what tasks can be completed.

This kind of facilitation of experience can be also understood through the concept of *affordance.* Donald Norman (1988) introduced the concept in HCI to mean “the relationship between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used.” In the context of interaction design this means that each feature of a certain system can be used in multiple ways, only parts of which might have been originally intended by the designer (ibid.).

Motivational affordance refers to properties that afford the user's motivation by answering psychological needs (Zhang 2008). In his paper, Zhang uses self-determination theory (STD, Ryan and Deci 2000b) to link the psychological needs of autonomy, competence and relatedness to design principles for ICT applications and provides examples of features that support fulfillment of those needs. Similar work has been since conducted in various contexts, for example persuasive technology (Weiser et al 2015).

Deterting (2011) extents the concept to *situational motivational affordance,* and reminds us about the importance of seeing experience as contextual and situated. When drawing parallels of features and motivation, it is necessary to keep in mind that certain features (for example challenges) can be experienced very differently based on their framing and interaction with other features.

The concept of motivational affordance has been widely accepted in UX research community. Besides autonomy, competence and relatedness in STD, there are a number of different types of needs that are suggested as useful in understanding user experience and affording motivation (Hassenzahl et al 2010). In this study I am

focusing on need frameworks based on STD. As I presented in chapter 3.1.3, it is one of the most applied motivational theories in HCI and games research, and also foundational to six needs suggested by Hassenzahl et al (2013) that are used in the Learning Engagement Evaluation framework.

3.3.1. Six needs suitable for experience design

Marc Hassenzahl's research on experience design uses need fulfillment for analyzing experience (2010, et al. 2010, et al. 2013). The research describes how an experience can be distilled into patterns, and how those patterns can be used to design new, pleasurable and meaningful experiences (2013).

Hassenzahl presents a model of 10 base needs from research of Sheldon et al. (2001, in Hassenzahl 2010). In the research by Sheldon et al. the participants were asked to report a satisfying life event, and then answer a set of questions about it. Based on their answers, Sheldon et al then classified the events based on the needs they fulfilled. The needs in the research were were *autonomy-independence, competence-effectance, relatedness-belonging, influence-popularity, pleasure-stimulation, security-control, physical thriving-bodily, self-actualizing-meaning, self-esteem-self-respect* and *money-luxury* (Sheldon et al. (2001).

Hassenzahl, Diefenbach and Göritz (2010) conducted further research on the applicability and saliency of these psychological needs in user experience research. Following the research setting similar to Sheldon et al (2001). The authors collected over 500 positive experiences with interactive products (e.g., mobile phones, computers) and by adapting the list of Sheldon et al. (2001) , classified them based on seven needs they identified as most relevant for user experience research: *competence, relatedness, popularity, stimulation, meaning, security and autonomy* (Hassenzahl et al. 2010).

The research found a clear relationship between need fulfilment and positive affect. Autonomy, competence and relatedness were especially noticeable in reported positive life events, followed by popularity, security and meaning. (Hassenzahl et al. 2010). In his later paper, Hassenzahl et al suggested six basic needs - meaning is excluded from the tested seven.

Table 1. Overview of a set of needs suitable for Experience Design (From Hassenzahl et al. 2013, based on Hassenzahl et al., 2010; Sheldon et al., 2001).

Autonomy	Feeling that you are the cause of your own actions rather than feeling that external forces or pressure are the cause of your action.
Competence	Feeling that you are very capable and effective in your actions rather than feeling incompetent or ineffective.
Relatedness	Feeling that you have regular intimate contact with people who care about you rather than feeling lonely and uncared for.
Popularity	Feeling that you are liked, respected, and have influence over others rather than feeling like a person whose advice or opinion nobody is interested in.
Stimulation	Feeling that you get plenty of enjoyment and pleasure rather than feeling bored and understimulated by life.
Security	Feeling safe and in control of your life rather than feeling uncertain and threatened by your circumstances.

Lallemand, Koenig and Gronier (2014) studied the validity of using six needs presented by Hassenzahl and colleagues (2013) in user experience assessment. Instead of using heuristics as such, they asked the 33 experts (from design, psychology, sciences, technology and HCI) to use UX cards to assess experience in using four different systems (e-commerce website Amazon, Facebook, Angry Birds -game, and a digital camera). The UX cards summarized the essence of each need, and the experts wereinstructed to identify elements that had a positive or negative impact on one or several needs within each assessed system. As a result, Both practitioners and researchers participating in the study found the UX Cards based on needs highly useful for both the design and the evaluation of interactive systems.

The final EAF Learning Engagement framework uses these needs as categories for heuristics. Although there are plenty of other propositions for needs suitable fo UX design, these six were found most salient and well suited; Autonomy, Competence and Reletedness are directly from STD, which is widely used for understanding engagement and experience both in the fields of education and user experience research; Popularity, Stimulation and Security were experimentally tested and validated by Sheldon et al. (2010) and Hassenzahl et al.(2010). As I will present later, the existing heuristic frameworks were easy to map to these categories, and our practical experience with product evaluations has given evidence that the categories cover the main aspects of different kinds of experiences.

In the next chapter I will discuss how the experience with the product can be assessed, and how engagement can stem from different kinds of positive experiences with the product

3.3.2. Positive experience as meaningful experiences

To understand what assessing user experience based on need fulfillment means, it is necessary to explain what is meant with a *positive experience*. Hassenzahl and colleagues (2010) argue that "the perceived qualities of an interactive product can be divided into instrumental, pragmatic and non-instrumental, self-referential, hedonic aspects ". *Pragmatic quality* refers to the product's perceived potential to support 'do-goals' - such as making a telephone call. It can be understood and evaluated through *utility* - how well the product supports performing necessary tasks, and *usability* - how easily and efficiently these tasks can be performed (Hassenzahl 2010). In their study on need fulfillment in experience, Hassenzahl et al (2010) suggest pragmatic quality being a ''hygiene factor": "enabling the fulfilment of needs through removing barriers but not being a source of positive experience in itself."

To understand engagement and what makes an experience positive, *hedonic quality* of use is more interesting. Hedonic quality can be judged in regards to a product's potential to support pleasure of use, and instead to do-goal, fulfilment of ''be-goals". With a "be-goal" Hassenzahl (2010, 13) refers to be-goals as ones that motivate actions

and make them meaningful. Making a phone call as a “do-goal” doesn’t have meaning in itself, but making a call to connect with another person fills a be-goal of “being related“, that gives meaning to be-goal.

Digital learning solutions differ from utility software in a fundamental way in that they don’t always want to lead the user to the correct place in the most effortless way. Instead, they want the user to make an effort to solve challenges. In this, interactive learning products are similar to games. In games, creating “fun” is different from making use of a system pleasant. Games do not fear giving their players challenges. They are not afraid of evoking negative emotions, such as frustration, anger and even fear. More likely, games can be defined through the challenge that they give to the player, or as Salen and Zimmerman (2004, 80) point out, game is an artificial conflict which is played to an outcome – win or lose. A game can also include what is called “abusive design” (Wilson & Sicart 2010). Game can be designed to be unfair, imbalanced and even deliberately causing the players strong negative emotions. According to Wilson and Sicart, abusive design should be understood as an aesthetic choice by the designer, an attempt to a dialogue between the game designer and the player through the act of gameplay.

In summary, the game designers need to balance the challenge and negative emotions the game might evoke with the positive emotions that arise from overcoming the challenge and experiencing the game world. The concept of flow is connected to good game experience (Salen and Zimmerman 2004, 336-337; Chen 2007,). In a flow state a person is highly focused and enjoys a task at hand. Flow is also defined by effortless involvement and high sense of control by the participant (Csikszentmihalyi 1991). The task at hand should not be too easy, because that will lead to boredom. Neither can it be too hard, since the participant needs to feel that they will be able to reach their goal.

Hassenzahl (2010, 48) describes a flow experience as *competence experience.* In this, he links flow to the basic psychological need of being competent - having the skills for overcoming challenges, meeting goals and having a sense of progression. When studying digital learning environments, it has been found that the state of flow has a positive correlation both to the students’ positive experience about the usage and to the learning outcomes (Esteban-Millat et al. 2014; Pearse et al. 2005). Keeping this in

mind, it might be difficult to find clear do-goals or pragmatic value for either games or many learning solutions. Although learning to program may have pragmatic value in the future, the task that the learner needs to perform in the interactive system would be meaningless or at least very frustrating without plenty of hedonic value.

Therefore it is not surprising that many learning products use game elements to motivate students. Gamification in education can be broadly understood as the use of game elements in non-game contexts, with the purpose of increasing student engagement and motivation. Various studies have reported significant correlations between gamified elements in the learning process and increased student motivations (Pechenkina et al, 2017, 2-3).

Motivational effects of gamification has been criticized in the basis that it relies on external motivation - that it is often reduced to points, badges and other elements, that can be considered rewarding only inside the designed system, but don't feed to the intrinsic motivation, participating in an activity for the enjoyment of the activity itself (Ryan and Deci, 2000a, 60), in a long term (Hung 2017, 1; van Roy and Zaman 2017). Research in psychology provides evidence that external rewards, such as money, can reduce intrinsic motivation, mainly because they decrease the person's feeling of autonomy and self-direction. This was summarized in a metastudy by Deci, Koestner, & Ryan (1999).

However, the research also suggests that if originally external motivational cues appeal to the psychological needs of the actor, these will lead to autonomous motivation (Deci and Ryan 2008a). This suggests that designed environments that use internal motivational mechanics that have no relevance outside it, for example a point system, can still create intrinsic motivation for the use, if these systems make the user feel competent, self-directed (autonomous) or having meaningful relationships through the system. To understand this in terms of Hassenzahl, performing an action in a game or learning environment, because it results in points, is a meaningless "do-goal", without pragmatic value, unless it is supported by hedonic value and "be-goals" - goal of being competent for example.

As mentioned, not all feelings in the experience of using learning solutions (or games)

are inherently pleasant. Also emotions that are usually considered negative, such as frustration in a face of a difficult task or sadness when encountering an emotionally moving narrative, can create powerful experiences that can be understood as positive. Therefore "positive" should be primarily understood as *worthwhile* and *meaningful (Hassenzahl 2010, 31)* - experience that has hedonic quality.

3.4. Evaluating design from the perspective of engagement

In the previous chapter I established that technology can create experience - or more specifically, the experience when using technology depends both on the affordances of technology and the context and goals of using it. When the right affordance meets the correct context, the person can also be inspired to use technology in a manner meant by the designer, and also in unexpected ways.

We also know that certain kinds of experiences can help learning and that it is possible to design for certain kinds of motivational affordances that can result in engagement. In this chapter I will discuss the ways in which the quality of different products can be evaluated, and especially how it is possible to assess the motivational affordances. The focus is on *evaluation of (user) experience* based on examining the product and analysing its design. I will shortly present different methods HCI uses to assess user experience, focusing on expert evaluation methods. The EAF Learning Engagement Evaluation is a heuristic evaluation that is done by experts first using the product and then checking it against a list of criteria set by the method. I will discuss the use of heuristics in general, and how the method is applied to assessment of learning applications and games , and especially what are the limitations of the method.

3.4.1. Heuristic evaluation in user experience research

As previously mentioned, In HCI the focus has traditionally been in assessing usability. Usability is a quality attribute that assesses how easy user interfaces are to use (Nielsen-Norman Group, retrieved 27.4.2020). There is an official ISO-9241-11 standard (Usability tests, 2020 retrieved 20.4.2020), which states that the product is usable when the use is effective, efficient and satisfying. The usability assessment methods can be divided to usability inspection methods and user-based methods. In

user-based methods the evaluation happens by involving real users, whose actions in the evaluated the system the usability expert is observing, and taking notes on the ways of use and problems the users are facing (Nielsen 1994a). In contrary to that, the usability Inspection methods involve the inspection of the interface by an evaluator. Since there's no need for designing and conducting time-consuming studies with users, the inspection methods are described as cheap, fast and easy to use and therefore labeled as "discount usability engineering methods" (Nielsen 1989).

Heuristic evaluation is one of the most used usability inspection methods. In heuristic evaluation the evaluator goes through the system and compares it against a set of design guidelines (heuristics). In usability research, one of the most foundational and well-known set of guidelines is 10 Usability heuristics by Nielsen and Molich (1990, Molich and Nielsen 1990), and later iterated by Nielsen (1994a), that were developed for expert evaluators to find usability problems.

A heuristic inspection is easy and fast to execute, but it should be noted that a single person evaluating the system will likely ignore some problems in the system (Nielsen, 1989, 1994). Therefore the method should be used by having several people independently go through the system, and then compare the set of problems they have detected. This is also true for the EAF Learning engagement method, where the evaluation is always conducted by a minimum of three evaluators. Nielsen (1994) states that three to five evaluators would be sufficient for reliable results, since adding more people would not significantly increase the number of usability errors found.

Since Nielsen's list of 10 usability heuristics, several other sets of heuristics have been developed. Mainly, there has been a need for a more domain specific set of heuristics that take into account the specific needs of the evaluated systems. For example, new heuristics have been developed to address specifics of new technology, such as VR (Murtza, Monroe and Youmans 2017) or touch screen devices (Inostroza et al. 2012). Sometimes heuristics are also developed for software that is used for a specialized purpose. In these cases, it is possible to formulate heuristics that refer to features that are known to be common in all software that are built for this purpose, and acknowledge that the list might not be applicable to all software (Rusu et al 2012).

Heuristic methods are mostly focusing on usability, and assume that the user experience designers wish to create positive emotions through good usability - they are removing obstacles of performing a wanted task. They wish to diminish the amount of anxiety the user feels when they interact with the system, and make the challenge of finding the right actions on the system as minimal as possible. When considering user experience, it is argued that this viewpoint is too narrow and misses out a wide variety of people's actual experiences and their motivations for using certain products (see for example Hassenzahl 2010, 45), and additionally, is insufficient; good usability and efficiency is simply a neutral baseline, that doesn't yet make the experience especially enjoyable or meaningful (Hassenzahl 2010, 27-28, also Arhippainen 2013)

There have also been some attempts to create generic user experience heuristics that could be applicable across different domains and purposes, but would still take into account a broader view on user experience design than Nielsen's heuristics. One such is suggested by Arhippainen (2013), and another,Lead, Follow and Get out of the way model was developed in IBM (LF&G, Kamper 2002). The LF&G model suggest that an optimal user experience is similar to good, facilitative learning relation -"Like a good teacher, mentor, or coach, the usable user interface leads the user to successful completion of tasks and goals; follows the user's progress and provides appropriate feedback and information when needed" (ibid).

3.4.2. Heuristic evaluation of learning solutions

The EAF Learning Engagement Evaluation framework is domain specific in a sense that it is created for assessing learning solutions. As presented in chapter 3.3.2., what is considered a pleasurable and meaningful user experience when playing a game is different from the user experience of a utility software.

In the context of learning technology, domain specific lists of heuristics have been developed for learning assessment systems (Sim et al 2006), that take into account the specific use of that type of software. Some heuristic frameworks have been also developed from pedagogical standpoints: For example, Squires and Preece (1999) propose an approach based on a set of heuristics for predictive evaluation of educational software that integrate usability and learning issues. Their heuristics consider for

example curriculum alignment and utility of the software in teaching practices.

As stated before, EAF Learning Engagement Evaluation framework is for assessing engagement and experience instead of pedagogical value. Therefore a closer comparison can be found from heuristics developed for assessing games and gamification, because these kinds of heuristics are more interested in what makes the use of system interesting and provides a good experience. I will present the frameworks that informed the development of EAF Learning Engagement in the chapter 4.4 when describing the creation process.

3.4.3. Limitations of Heuristics User Experience assessment

As presented in previous chapters, assessing an experience created by technology always has limitations; experience is situated, subjective and context dependent. This is commonly recognized by creators of heuristics and design guidelines. In the case of Gameful design heuristics, the authors state that "[T]his method cannot evaluate the user experience; its goal is to evaluate the system's potential to afford a gameful, engaging experience. " and its possibilities in doing so greatly depends on the experience and expertise of the evaluator (Tondello et al. 2019).

However, research shows that a considerable amount of issues identified by an expert evaluator are actually based on his expertise and personal judgment, and not on the set of heuristics that are used in the assessment, and therefore the result is similar to less format evaluation, where the expert doesn't use a list of heuristic (Cockton and Woolrych 2001). In Learning Engagement evaluation this is also a possibility, but as I will present later, the method requires the evaluator to consider each heuristic. In this, it is actually closer to a survey than heuristic evaluation as it is commonly conducted.

The main limitations of heuristic evaluation result from expert variability - different evaluators finding different sets of problems - and the fact that evaluators tend to overreport problems that aren't necessarily problems (Lallemand, Koenig and Gronier 2014). Therefore heuristic methods are recommended to use in combination with user-based methods like user testing (Hvannberg, Law and Larusdottir 2007).

In the study of Lallemand, Koenig and Gronier (2014), where need based UX cards were used for expert evaluation, experts tended to link elements of evaluation to positive needs rather than to negative ones. The authors state that this may be a major difference in assessment of UX versus usability. This should be also kept in mind when interpreting the results from Learning Engagement Evaluation.

4 THE LEARNING ENGAGEMENT EVALUATION FRAMEWORK

In literature review I presented how engagement can be perceived in education, and how it is linked to motivation through need fulfillment. I also discussed how need fulfillment can be used to understand motivational affordances and user experience of (learning) technology, and presented heuristic evaluation as a method for assessing it. In this part of the thesis I will present he EAF Learning Engagement Evaluation method, and how the set of heuristics was created, tested and validated. I will document the steps that were taken in the iterative process and link the final model to existing research. Finally, I'm describing the software that was developed in EAF and how it is used by UX experts and teachers to conduct quality evaluations with the Learning Engagement Evaluation method.

4.1. Definition of "learning solutions"

Before going in to the evaluation method itself, it is necessary to define the object of the evaluation method. The EAF Learning Engagement Evaluation is developed for assessing learning solutions and educational tools. As defined in the beginning of the chapter 3.2., A "learning solution", in this study can be used as a synonym for educational technology or learning technology. However, as was the case in European Commission survey (2019b), "educational technology" can be also used as an umbrella term that covers all technology used in education, including devices like tablet computers or video projectors

Therefore it is necessary to define set a criteria for learning solutions this research is interested in and which can be evaluated with the method:
[1] They are meant to be used by learners [2] They have some types of interactions in them that are meant to be engaged with for learning purposes [3] The solution is aiming to teach either a skill or a knowledge goal.

The first criterion excludes technology that is used solely by teachers, such as administration tools. It would be possible to assess what kind of positive experiences they provide for teachers, but as these kinds of solutions mainly have mainly pragmatic value, the evaluation framework focusing on hedonic value and emotions isn't entirely relevant. The second criterion emphasizes interactivity and therefore excludes non-interactive learning solutions, such as books. The last criterion is important, because evaluating general hardware that is used in classrooms, such as tablet computers or video projectors, or general utility software, (word processors, communication software), would be difficult to assess in a meaningful way. These also have mostly pragmatic value. However, the line between an utility software and a learning software can be fuzzy; a software developed for eg. video editing can include elements that actively aim to teach certain skill goals, for example related to visual storytelling or sharing of information.

4.2. Goals for the Learning Engagement Evaluation method

The Learning Engagement framework was created to complement the pedagogical analysis in the EAF learning solution evaluations. The need for a separate section covering user experience rose from the fact that the pedagogical analysis did not have a systematic way of identifying potential usability or user experience problems in the products and formulating them for the client. For example, insufficient feedback from a task can be a pedagogical problem, but a poorly designed system error message might not be, although it does affect the learner's experience.

The pedagogical analysis is done by experts and doesn't involve interacting with real users (learners) of the product. Therefore it was necessary to select a usability inspection method that could be also used by an expert without involvement of users. When considering the choice of the method, *cognitive walkthrough method* (Wharton, et al. 1994) was also tried. When using the method the expert evaluator defines, what are the most common tasks the user wants to perform when using the product, and then step by step tries to identify factors that either help or prevent performing the task (ibid.) This turned out to be too usability focused to meet the second objective for the analysis; analyzing the positive feelings the product evokes.

The analysis framework should help to cover not only the user experience problems, but also what makes the system interesting, engaging and fun. Identifying both the strengths and development areas of the product was part of the pedagogical analysis, and we saw that the Learning Engagement framework should do the same in order to be more than a simple usability analysis of the product. To make reporting of the results systematic and concrete, we wanted to categorize the heuristics under themes that could be used as titles for different aspects of analysis. In the end, the need fulfillment framework by Hassenzahl et al. (2013) was selected as the underlying theoretical framework and the psychological needs in it were adapted to create a classification for the Learning Engagement Evaluation method.

4.3. Method of creating the set of heuristics

Creation of heuristics doesn't have a clear cut, formal method that would be generally accepted and applied through the HCI research community. Rather, creation of new heuristics or adaption of existing ones tends to be based on the developer's experience (Rusu et al. 2018). In a metastudy examining creation of domain specific new heuristics, Hermawati and Lawson (2016) conclude that most of the researchers follow two steps; 1) extraction of information, and 2) transformation of the extracted information into heuristics. However, there is great variation on how these steps are followed, and even more ambiguity in validating the created heuristics.

4.3.1. Creation and validation methods used in related heuristic lists

The research that most informed the final list of heuristics also used several different ways of creation and validation. All of them based their work on existing literature and heuristics, but have various ways of using the research and validating their heuristics. I chose both gameplay heuristics and other user experience heuristics as a basis for my own work.

In their heuristics for Gameful Design Tondello et al (2016). used STD as a guiding theory for mapping existing heuristics frameworks (most of which also used STD) and developed their set based on them. Use of STD provided a theoretical foundation and validation for the work. The actual heuristics were tested in a later experimental study

(Tondello et al 2019). Because of their close link to STD, Gameful Design heuristics ended up having a big influence to the final set of learning engagement heuristics.

Korhonen and Koivisto (2006) tested their initial set of heuristics for mobile touch screen games by evaluating one game themselves. After re-visiting the list they organized a review session with experienced game designers. When the final list was established, it was used to evaluate five not yet published mobile games by two to four expert evaluators each. Their heuristics were practical and referred to game elements in a concrete manner. Therefore referencing them especially early on the process gave valuable insight to various factors and aspects of use.

Heuristic Evaluation for Playability (HEP) by Desurvier et al. (2004) was a rigorous list and the heuristics had reviewed by several playability experts and game designers. The list was also tested by comparing the findings of experts using the heuristics to interview comments made by users who had played the evaluated games.

Inostroza et al.(2012) used a formal method proposed by Rusu et al. (2012) for developing and validating their heuristics. This was the only formal method I discovered on the topic, and I found it very useful in my own process.

4.3.2. Methodology for creation and validation of heuristics

The in developing learning engagement heuristics started in 2017. At that point, I didn't follow a set methodology, but rather started experimenting with different existing heuristics and validating my own through basing them on literature and using tests and validations similar to those I found from the research of the source heuristics. The methodology presented by Rusu et al. (2012) has the benefit of clearly defined steps to follow and strong emphasis on validation of the heuristic. Therefore I am documenting my journey by comparing it to their methodology of creating heuristics. This should bring transparency to the process and also show the potential needs for more validation. Although the methodology is proposed for developing new domain specific heuristics, and my aim was to create general heuristics (or at least the domain of "educational solutions" is a very broad one), I see that the method still has very important points and

will be valid for most part. Next, I will present the steps and shortly describe the work related to each step.

Methodology for creation and validation of heuristics (Rusu et al. 2012)

The aim of the three first steps is to collect bibliography and highlight the most important characteristics of the previously collected information, and to formalize the main concepts associated with the research.

Step 1: Exploratory stage: *perform a literature review.*

Step 2: Experimental stage: *analyze data that are obtained in different experiments to collect additional information that has not been identified in the previous stage.*

Step 3: Descriptive stage: *select and prioritize the most important topics of all information that was collected in the previous stages.*

When starting my work, I did plenty of research on the topic. This work is documented in the literature review of this thesis.

Rusu et al. also point out that in step 2, information can be gathered from other sources besides research literature, such as making experiments or analyzing experiments done by other researchers. This is in line with the findings or Hermawati and Lawson (2016), who in their analysis found that information extraction can happen in four different ways; 1) using one or more theories as a basis to identify aspects that were relevant for users' interactions; 2) studying the context of use and identifying aspects that were relevant for users, 3) studying and synthesising reported pertinent usability issues and/or existing heuristics/guidelines, and 4) developing a corpus of usability issues and identifying pertinent issues.

Hermawati and Lawson (ibid) recommend that researchers consider all of the mentioned options and possibly use a combination of them. In my work I relied on literary sources for existing heuristics and qualities of user experience.

Step 4: Correlational stage: *match the features of the specific application domain with the usability/UX attributes and existing heuristics*

In this step, the idea was to check how well the existing heuristics match the need. When developing domain-specific heuristics, for example heuristics for learning management systems (LMS's), the researcher should compare the heuristics to the commonly found features in that domain. The method also recommends grouping the heuristics for finding common themes and in that way reducing its complexity.

In my work, the need was to create a framework that would holistically cover the main parts of the user experience in a variety of different types of learning solutions. Early on it became clear that analyzing the spectrum of "all learning solutions" to find common features in them would have been difficult. From the beginning it was clear that because EAF evaluation already included a method for assessing pedagogical quality, I would exclude many aspects related to pedagogical design of learning solutions and focus on more universal engagement factors.

When starting the work in 2016 I had four years of industry experience as a game- and educational software designer. This greatly impacted my view on what different kinds of learning solutions have in common and what kinds of features make them engaging, and this helped me to collect a broad list of existing heuristics and start evaluating their usefulness. After some experimentation, I made the decision to use Hassenzahl's et al.'s (2013) framework of need fulfillment as a theoretical backbone of my own list and used it to guide the selection and adaptation of heuristics in a very similar manner to what is proposed in the method.

Step 5: Selection stage: *keep, adapt and/or discard the existing sets of usability/UX heuristics that were selected in Step 3*

Step 6: Specification stage: *formally specify the new set of usability/UX heuristics.*

After Steps 5 and 6 the researchers should have a concise, clearly written set of heuristics. In stage six, Rusu et al. (date) propose a template for formally writing a heuristic. My final framework that was formed in 2018 included many qualities of the proposed template, and that will be explained in 4.5.

Step 7: Validation stage: validate the set of heuristics through several experiments in terms of their effectiveness and efficiency in evaluating the specific application. The method proposes validation through heuristics evaluation, expert judgement and user testing. I will discuss this in chapter 4.4.3.

4.4. Creation process

In this chapter I am going through the early iterations of the framework to bring visibility to the creation process. First, I am describing the first attempt on creating categories for existing heuristics that I found relevant and interesting, and then present the second iteration, that was piloted in product evaluations made in EAF during 2016 and 2017.

4.4.1. Existing heuristics and first iteration of themes

In 2016, the first step of creation was looking for the reference literature on heuristic evaluations. The main requirement for the heuristics were that they covered multiple aspects of user experiences, and weren't too tied to a certain platform or media. Game design and gamification heuristics had special relevance for this research. as presented in the literature review, interesting and engaging game experiences have similar qualities to engaging learning experiences. Also, based on the experience we at EAF had, many of the learning solutions we saw had a gamified approach or game elements. I was also familiar with playability heuristics due to my background as a game designer.

In the first iteration of the research I selected three existing heuristic frameworks for a closer study; Korhonen and Koivisto's (2006) 'Playability Heuristics for Mobile Games' has influenced several new sets of heuristics (Paavilainen 2010; Pinelle et al. 2009), and the heuristics were short and despite being made for specific domain, quite generally written; 'Heuristic Evaluation for Playability (HEP)' by Desurvier et al. (2004) is a rigorous list and the heuristics have been empirically tested. Although a second iteration of the heuristics has been developed and the old list has also been expanded and refined to cover specific game genres (Desurvire and Wiberg 2009), I decided to use the original HEP list, because it had several heuristics about game story and a categorization that was somewhat similar to Korhonen and Koivisto's

heuristics. The third framework was ‘Gameful Design Heuristics’ by Tondello et al. (2016) that was selected because it focused on motivations for use and was developed to also suit gamified applications; not only pure games. This was well in line with our aims for the heuristics.

Because we wanted to have clear categories in the analysis, the first themes I tried were an adaptation of the Jacob Nielsen’s 10 usability heuristics (Nielsen 1994a, see the Table 2). The reason for selecting Nielsen’s heuristics was that they were well-recognized and widely-used, so I felt they would provide a very credible backbone to my work, if they could be applied. Nielsen’s heuristics had also influenced both Korhonen and Koivisto’s Playability Heuristics for Mobile Games and HEP heuristics, so I assumed they would work well together. I analyzed the selected heuristic frameworks and matched the heuristics in them to the themes I had derived from Nielsen’s list. Matching was quite robust and intuitive, because the aim was just to test if there were related heuristics for the themes and if some new themes would emerge from heuristics that did not fit.

Table 2. Mapping of user experience and gameplay heuristics to Nielsen’s 10 Usability heuristics

Theme	**10 Usability heuristics**	**Heuristic Evaluation for Gameful Design**	**Playability Heuristics for Mobile Games**	**Heuristic Evaluation for Playability (HEP)**
Visibility	#1: Visibility of system status	C2, C3, C1, I1, I2 , I8, I6, I7	GU3, GU4, GU9, GU11, GP1	GP3, GP13, M1, M2, M3, U1, U5, U6,
Familiarity	#2: Match between system and the real world	E1, I15	GU5, GU11, GP7	GP2, GS2, GS7, U12
Control and Freedom	#3: User control and freedom	C6, I6, I5, I9, I8, I10, E3 , E1	GP1, GP4, GP9, GP10, GU3, MO1,	GP10 GP12, GS5, U2
Consistency	#4: Consistency and standards		GP12, GP13, GU3, GU7, GU6,	GP2, GP7, M1, M4,M5,M6, GS1

Recognition	#6: Recognition rather than recal	C3, I3	GP2, GP13, GU1, GU5, GU11	GP4, GP7, U8, U4, U12
Flexibility	#7: Flexibility and efficiency of use	I8	GP8, GP10, GP11, GU2, GU3, GU8, GU11	GP6, M6, M7, U6, U10
Asthetics	#8: Aesthetic and minimalist design		GU1, GU2, GU3	GP9, GP11, U7, U9, U10, U12
Errors	#9: Help users recognize, diagnose, and recover from errors	C6	GP14, GU10, GU12, MO3	
Helps	#10: Help and documentation	C6	GU12	U6

Table 2 shows the codes of heuristics in the selected frameworks mapped to Nielsen's 10 Usability heuristics.The result of the analysis was that the themes from Nielsen's heuristics were insufficient. When doing the mapping, it became clear that especially Gameful Design Heuristics (Tondello et al 2016) didn't fit those themes well - many heuristics could not be attached to any of the themes or the attachment was artificial. In particular, heuristics that referred to game goals, challenge, or story could not find a natural place. Secondly, some categories like Visibility, Control and Freedom, and Consistency had a lot of very different kinds of heuristics mapped to them.

For example, I categorized GP1: "The game provides clear goals or supports player created goals" (Playability Heuristics for Mobile Games, Korhonen and Koivisto 2006) to match both Visibility (the user should be clearly aware of what the goals are) and User Control and Freedom (users can create goals themselves). However, neither of these categories are a clear fit, and as a theme Visibility (of the system status) is related to several other aspects, such as progression indicators and even story progression. All of these: goals, progress, story, are very important themes on their own, and feel downplayed or even hidden, if grouped together.

This should not have been surprising, because many studies in games and playability note that traditional usability evaluation methods and heuristics don't cover all relevant aspects of games well (see for example, Sánchez 2012).

For my work, the most important finding from this short exercise was that the themes should reflect the reasons why someone would use the learning solution and want to continue using it. With usability heuristics, such as Nielsen's list, the goals of use are interesting only in a sense of how all the other aspects support fulfilling these goals - they don't take a stance on whether these goals are meaningful or not.

4.4.2. Second iteration and psychological need fulfilment as themes

For the second iteration, I decided to use the need fulfilment model of experience design proposed by Hassenzahl et al. (2013) for categorizing the heuristics and developing suitable themes. The framework is based on 10 psychological needs definitions from the research of Sheldon et al. (2001), and Hassenzahl has narrowed them to the six most significant ones for user experience design.Sheldon et al. (2001) used SDT as one of the foundational motivational theories in their study. Since SDT has been successfully applied to understanding motivational affordances in ICT (Zhang, 2008) and in games and gamification (Deterding 2011, Weiser et al. 2015, Tondello et atll 2016; 2019), I felt confident in using the need framework.

I used the same set of heuristics to match them with the need fulfilment model, and this turned out to work well. The Gameful Design Heuristics by Tondello et al. (2016) used similar categories as Hassenzahl (2013): *Autonomy and Creativity (Autonomy), Challenge and Competence (Competence), and Relatedness.* These "big three" are derived from self-determination theory (Ryan and Deci, 2000), and in the experimental research of Sheldon et al. (2001), were found to be the most important of the 10 they proposed.

Based on this matching, I wrote an initial set of heuristics. For this first version, I used

game terminology, and referred to the users as players and products as games. The reference heuristics for my list were game and playability heuristics and we had decided to pilot the framework with game products. In addition to heuristics, I also created help questions that would summarize important qualities of the experience and explain the focus of the heuristics in each category better.

The category descriptions were taken from descriptions by Hassenzahl et al. (2013). I made small adaptations to them to make them better fit game context. For example, the original description for Autonomy (ibid): *"Feeling that you are the cause of your own actions rather than feeling that external forces or pressure are the cause of your action",* was changed to: "*Feeling that your actions in the game are based on your own decisions rather than feeling external pressure to choose a certain action".*

A major change to the original need fullment framework was to change "Popularity" to "Respect". Popularity as a need was taken to Hassenzahl's framework from the psychological research by Sheldon et al. (2001), where its qualities were related to social influence and dominance. This was slightly problematic in the context of assessing the user experience of designed products, where the interaction can happen only between the product and the user. Hassenzahl (2013) gives an example of a person helping someone else and in that way fulfilling the popularity need. In a digital context, these kinds of interactions can happen, but that would require multiplayer options or other multi-user mechanics. I decided to switch the viewpoint of popularity to *respect*, specifically, how the product designer has shown that they know their users and respect them by taking into account different types of users - novices, the more experienced / experts, users of certain genders, backgrounds, and capabilities. This would be in line with the original meaning of the need of popularity, since showing respect to someone results in the feeling of being popular (Sheldon et al., 2001)

Table 3: The Learning Engagement framework - the first version (June 2016)

Help questions	**Heuristics**
Autonomy: *Feeling that your actions in the game are based on your own decisions rather than feeling external pressure to choose a certain action.*	

Do I care about the thing the game wants me to do?	The game supports meaningful goals
Can I make choices of what I want to achieve on the game?	The game supports self-made goals
Do I have a chance to express myself?	The game gives you enough information to make meaningful choices.
Do I have things to do in the game even when I'm lacking company?	The game supports different playing styles
	The players can express themselves
	The design overcomes a lack of players and enables soloing.
Competence: *Feeling that you are capable and effective in your actions rather than feeling incompetent or ineffective.*	
Do I feel like I can learn the skills necessary for the game?	The game has a challenge that is optimal
	The game rewards you in a meaningful way and according to the challenge
	The first time experience is encouraging
Do I feel like I have a chance of winning the game?	The game should cater for both inexperienced and experienced users. Players can eg. skip tutorials or choose wanted difficulty levels
Can I feel successful and proud of myself when playing?	The game doesn't offer "cheats" or shortcuts that make tackling the game challenges irrelevant (these are usually game design faults)
	The point/score system measures the thing that should be on the core of the challenge.. Eg. math skills or vocabulary instead of eg. speed, memory, pattern recognition etc.
Relatedness: *Feeling that you have contact with people who care about you rather than feeling lonely and uncared for. In the game you also feel connection with fictional characters and events in it.*	
Do I feel like the other people in the game want me to communicate with them on the game?	The game helps the player to find other players and game instances.
	The game supports groups and communities.
	The game supports communication & there are good reasons to communicate

Do I feel like the game wants me to play it and invites me back?	The game notices, when you are away for a long time and shows it “cares” about it.
	The game rewards for logging in regularly
Do I care, what happens to other characters or people in the game?	The push notifications are meaningful and make the player want to return to the game.
	There is a good reason to communicate with others
Respect: *Feeling that you are liked and respected, and the game takes you into account as a player rather than feeling like a person whose opinion nobody is interested in.*	
Can I get meaningful feedback from the game and other players?	The game gives you feedback on your actions
Does the game make me feel that I am a wanted player?	The game should cater for both inexperienced and experienced users. Players can eg. skip tutorials or choose wanted difficulty levels
Is the game discriminative/does it provide a narrative that is offencive?	The game addresses the player in a way that doesn’t make assumptions on player’s age, gender, race or origin. (Note: only direct addressing of the player, addressing the player character is different)
	The game doesn’t enforce discriminative narrative
Stimulation: *Feeling that you get plenty of enjoyment and pleasure rather than feeling bored and understimulated by the game*	
Are the challenges in the game too easy?	There are no repetitive or boring tasks
Do I need to repeat things I have learned unnecessarily?	The game has a challenge that is optimal for you
Am I intrigued of what will happen next in the game?	The game graphics, sounds and other elements support the story and gameplay in a meaningful way and are pleasant.
Do I find the game visually enjoyable and rich?	
Security: *Feeling that the game is a safe environment for having fun and trying out things rather than feeling uncertain of the consequences or threatened by other players*	
Does losing make me feel so angry that I would rather not have played at all?	The design minimizes deviant behavior. The game has a way to report and possibly block toxic players.
	The player does not lose any hard-won possessions
Am I scared/offended of things other people are able to do to me in the game?	The player cannot make irreversible errors. Points that lead to losing the game without a chance for a retry without a considerable effort should not be possible

	Is making errors beneficial or punishing? Do you learn something from each error you make?

The first iteration of the learning engagement framework can be found from the Table 3 The Learning Engagement framework - the first version (June 2016). I will describe the aspects and heuristics more closely when presenting the final framework.

4.4.3. Validation of the model

Once the first version of the framework was created, it needed to be tested and validated. In step 7 of Rusu et al (2012). method for creating heuristics proposes validation through heuristics evaluation, expert judgement and user tests. In heuristic evaluation test, the same product is evaluated with the tested set of heuristics and with another set of control heuristics to find out, which one results in discovery of more valid problems. In a user test validation, the product would have been evaluated by an expert using the heuristics, and the results would be compared to problems found in actual user testing. Both of these methods focus on finding problems, whereas the Learning Engagement framework should work also as an analysis tool for identifying the factors resulting in positive experience. Tondello et al.'s (2016) heuristics for gameful design could have worked as control heuristics, because they aim to holistically assess the experience and put emphasis on fun and meaningfulness. However, they are more focused on game systems than the Learning Engagement framework and therefore the validation with them should be done with (educational) game products.

Using a user test for validation might have been possible by asking the users what makes the use of a system fun and engaging or frustrating and problematic and comparing that to the results of the expert evaluation. However, this was not done due to resourcing reasons. We also saw that using the framework ourselves to analyze products and then discussing the findings with the development team gave us indirect feedback about the users - the developers often either confirmed our findings by telling that they matched the feedback they got from the users or sometimes questioned it. However, this was not formally documented.

The framework was validated by using expert judgement. First, it was used to evaluate

three learning solutions; An educational mobile game for learning English, a gamified social feedback app and a quiz creation tool. This was done by three people, myself and two other EAF experts.

The results of each evaluation were reported in an evaluation report that was delivered to the client. The heuristic categories were titled “the six aspects of learning engagement”, and each of them had a section on the report. Although the evaluation was done against the list of heuristics, the results from the evaluation used the help questions to summarize the findings. There is an example of this in Image 1. The evaluation used a numeric scale from 1-3 to describe the comments:

3 = **Well supported:** There are several well executed features which support this aspect of engagement.
2 = **Supported:** The game takes into account this aspect. Some improvements could be made in order to make the support better.
1 = **Not Supported:**There are issues with the engagement in this area.

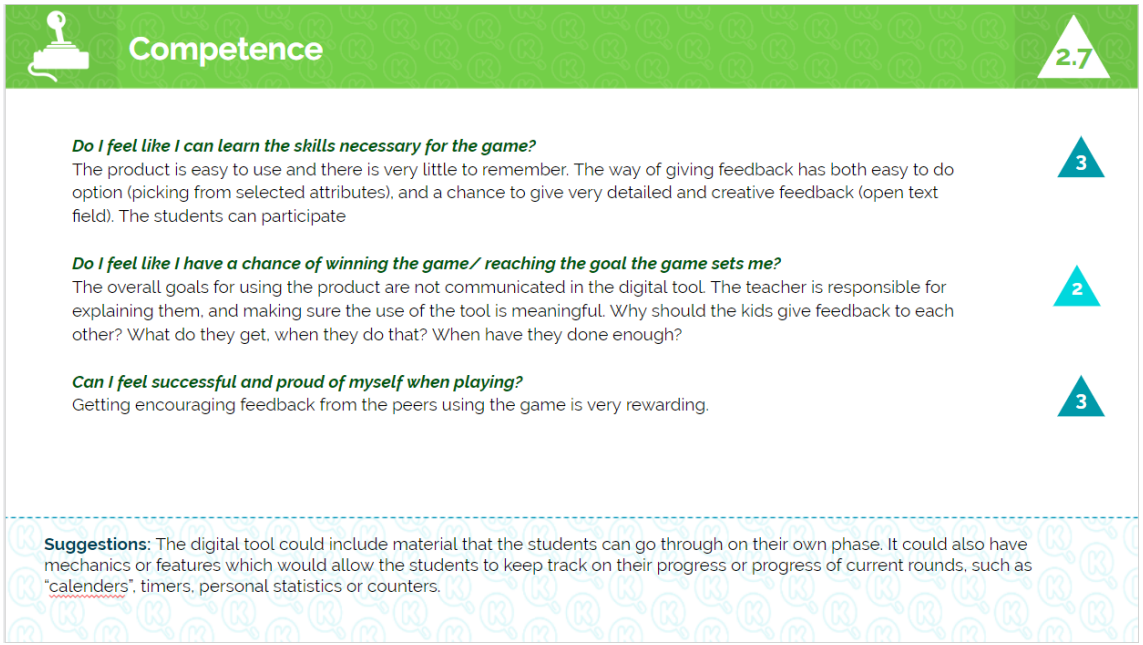


Image 1: A page of the Learning Engagement evaluation report made for a client, Education Alliance Finland, 2016.

The pilot tests provided positive results and showed that 1) the selected themes worked well in providing a holistic view of the experience, although the clients needed an explanation for understanding the terminology, 2) the heuristics and help questions both were relevant in the assessment, and the questions were understandable for the clients, and 3) the clients especially appreciated the suggestions by the evaluators.

During 2016 and early 2017 the framework was further tested and iteratively developed by evaluating 21 other products that sought EAF's consultation. These included digital learning games, board and card games, content creation solutions (for example video editing), student assessment systems and even a course program that mixed face-to-face teaching and the use of a digital learning solution. The results from this work were again positive; the framework was used by four different experts and it proved to be suitable for a great majority of products. On a small number of products it was necessary to admit that the framework was not suitable. Firstly, some learning management systems did not have enough student interactions that could be assessed for engagement factors. These products were mainly targeted to teachers and school administrators, and offered great pragmatic value for them. Secondly, some products were in their very early stages, where it was difficult to assess the concept holistically. It was possible to conduct the learning engagement analysis based on game design documents, but the results were not meaningful enough to the team receiving the feedback. Some aspects, such as Stimulation, Safety, and Respect were especially difficult to assess simply based on the concept design.

During the testing period, the main need for a change was to make the framework suit other apps besides games, which meant reformulating the heuristics. In many instances it was not problematic to simply change "game" to "product" but some texts required more work. The second change was to bring in heuristics that took into account specific features of learning apps. For example, the heuristic *"Progression in the product depends on succeeding on things relevant for learning."* was added to Competence. This was based on our EAF team's experience with learning games. Quite often we encountered a product, where a user was required to play a game that needed some motor or coordination skills, such as controlling a player character in a platformer game. The actual learning objective could be to, for example, teach basic math and the user would need to jump to a platform with the correct answer to a math task. However, this kind of design could become problematic if the reward system in the game was based

on the time the user took to find the correct platform. In that case, the challenge was more related to jumping correctly in a good sequence than learning to solve math problems. There were other changes such as this, and I will discuss them more when presenting the final model.

After our own tests and iteration we at EAF wanted to have an external expert validation of the framework. In June 2017 the framework was evaluated by Janne Paavilainen, an expert games researcher specializing in playability and heuristic evaluation. His main feedback was that the categorization based on need fulfillment and the user heuristics were theoretically well-justified and both the help questions and heuristics were clearly written and easy to use. His main recommendation was to streamline the list of heuristics - there was some repetition and many of the heuristics were quite focused and specific so they could be potentially ill-suited for a wider spectrum of products.

4.5. The final model

Based on the feedback and test evaluations, the final version of the learning engagement framework was created. To answer the critique and make the framework more streamlined, I decided to include some of the help questions as heuristics and leave out some of the more specific heuristics. Also, new heuristics with a focus on learning solution design were added.

Table 4: The Learning Engagement framework - the final version October 2017

	Heuristic	**Explanation**
	Autonomy: *The users actions in the product are based on their own decisions rather than feeling external pressure to choose a certain action.*	
A1	The user can create their own goals for the use.	*Does the product have enough features to enable doing a variety of interesting things? If the product is linear and closed, can the user choose at what pace to use the product or what parts of the product to interact with? Can the user choose an approach or strategy to solve problems?*

A2	The product motivates the use well.	*How does the product show why the skills it teaches are important or why its content is interesting? How does the product reward use and how well does it draw you in ?*
A3	It is easy to understand, what is the goal in using the product.	*The purposes and reasons of use are clear. How well does the product tell you what it can be used for and what can you learn from it?*
A4	The product sets limitations for using it when and where I want to, and the limitations feel unnecessary or annoying.	*The product imposes unnecessary restrictions on use. Are there time limits or mechanics in the product that prevent use when I want to? Is access to some content restricted? If so, are the restrictions justified?*
A5	It is possible to make choices, and the different choices have clearly different and meaningful outcomes.	*Can the user see if he is making the right or wrong choice? Are there any narrative choices in the product? If a product is an open ended tool or simulation, is it possible to do a wide variety of things, or is the range of tools too narrow?*
A6	It is possible to use creativity and express yourself when using the product.	
Competence: Feeling that you are capable and effective in your actions rather than feeling incompetent or ineffective.		
C1	The product rewards the user in a meaningful way and according to the challenge.	*Is there an adequate and frequent reward for success? Is the user progressing fast enough or does the product require disproportionate effort for progress?*
C2	The product gives you enough information to use it efficiently.	*The product provides sufficient information for effective and meaningful use. Is there enough instruction and tutorials in the product?*
C3	Navigation in the product is easy and intuitive.	*Do I understand what all the parts in the product do? Can I easily find what I want and get back to the previous state?*
C4	The challenges and tasks in the product feel optimal for the targeted users.	*Is the difficulty level of the product consistent and appropriate for the target group? Is it possible to solve the most difficult tasks with skills that can be learned and practiced within the product? Is the user interface suitable and usable for the target audience?*
C6	Progression on the product depends on succeeding on things relevant for learning.	*Is progress and success dependent on thinks that the product is teaching to the user, or is it possible to be successful by being good at a skill that is not essential*

		for learning? For example, a language learning game where good scores require good reaction speed rather than broad vocabulary?
C7	The first time experience is encouraging and it is easy to learn to use the product.	*Is the interface easy to understand or learn? Is there a tutorial in the product? Do the tasks at the start of the use present easy challenges that create feelings of success? Can a novice user easily find help or, for example, ease the level of difficulty?*
C8	Experienced and advanced users can find more challenge in the product.	*Does the product offer more difficult tasks for more advanced users? Can creativity and skills be used to do more complex things? Can I adjust the difficulty levels of the product myself?*
C9	It is possible to feel successful and proud of myself when I am using the product.	*Does the product tell you when your are doing things right? Can I always solve more and more difficult problems when I am progressing in the use? Is it possible to create things with the product that make me proud of myself and my skills?*
Relatedness: *The product supports meaningful contact with people who about your actions rather than feeling that the contact is one-sided or meaningless. The user can feel connection with fictional characters and events in the product.*		
RL1	The story or fictional world present in the product motivates learning.	*Is the fictional word immersive and interesting? Does it provide a reason for exploring the learning activities further?*
RL2	The product uses language which makes you feel welcome and cared for.	*Are the texts and other messages clear and polite to the user? Do they have humor? Is the system addressing the user at the right moment? If the user makes a mistake, does the product encourage them to try again?*
RL3	The visuals and characters in the product are suitable for targeted users.	*Does the product look interesting and well-polished? Are there any interesting fictional or real characters in the product?*
RL4	The product supports social interaction, such as multiplay or sharing of content with other people.	

RL5	The product provides examples of or motivation to learn the skill it tries to teach.	*Does the product tell you why the skill or knowledge it teaches is important? Does it provide examples of what you can do with the skill in the real world, or how the information provided by a product relates to real life? For example, does the product have links to other users' creations or other inspiring content?*
RL6	The product supports communication with other people and there is are good reasons to communicate	*Does the product have direct communication, such as collaboration or competition? Is this implemented in a way that makes sense and is fun for everyone?*
Respect: *Feeling that the product takes the user into account as a capable and desired actor rather than feeling that the user's opinions and experiences are neglected.*		
RS1	The product gives clear feedback on all your actions	*Is it clear what each available function does? If you make a mistake or take a correct action, is the feedback provided clear? Do the error messages indicate why the operation failed?*
RS2	The product doesn't make assumptions on player's age, gender, race or origin.	*If the product has not been made for a specific target group (eg. Finnish first-graders, Kenyan girls), does the product take into account all potential users? Does the product refer to the user in eg. gender-related terms or terms that assume certain life style?*
RS3	The product doesn't include discriminative narrative or enforce unnecessary stereotypes.	*The product should not contain discriminatory or stereotypical language or imagery. If the product has characters, what kind of characters are shown? What are the main characters and secondary characters like? If the user can choose an avatar for themselves, is there a wide variety of interesting characters in the selection?*
RS4	The product is suitable for both inexperienced and experienced users. Players can eg. skip tutorials or choose wanted difficulty levels.	*The product should cater for both beginners and advanced users. For example, the user can skip the tutorial modes or select the desired level of difficulty. If the product is used for a longer period of time, does the product take it into account?*
RS5	The product doesn't have bugs which cause errors or crashing.	

		Does the product work well with tested equipment? Are layouts working with each supported screen relation and size?
Stimulation: *Feeling that you get plenty of enjoyment and pleasure rather than feeling bored and understimulated by the product.*		
ST1	The product encourages exploring it further.	Is there a good amount of content and features the use can explore? Does the product suggest new things to do?
ST2	The product's challenge level is optimal for the targeted users, or it can be chosen.	*Are the challenges in the product difficult enough to keep the user interested? Are they too challenging and lead to frustration?*
ST3	The user doesn't unnecessarily need to repeat things which they have already learned.	*Does the product stagnate? Does it require grinding or repeating the same things over and over again until the user is bored.*
ST4	The product's graphics, sounds and other elements support the narrative and user experience in a meaningful way and are pleasant.	*Are the most important things in the interface highlighted? Is there feedback from sounds? Is the overall visual looks and soundscape suitable and appealing for the target group and does it fit to the purpose and mood of the product?*
Safety: *Feeling that the product is a safe environment for having fun and trying out things rather than feeling uncertain of the consequences or threatened by other users*		
SA1	Making errors is beneficial. Everytime you make an error, you learn something from it.	*Can the user try out things in a way that is encouraging? If they make a wrong action, does the product notify them about it*
SA2	There is a way to report and possibly block misbehaving users.	*If there are social features, is it possible easily report misbehaviour to the developer or to the teacher monitoring the situation? Is it possible to change settings to block people who are misbehaving?*
SA3	The product doesn't include content or advertising which would be harmful for the targeted users	*Is the content suitable for targeted users? Is there links to outside content that might be harmful?*
SA4	If the user shares content - their work, their comments or anything else - it is always clear,	*The user should always be aware, who can see, comment or edit any inputs the user makes. Is this clear for user profiles, creations and everything else?*

	who has access to the shared content.	
SA5	The user cannot make irreversible errors. Points that lead to restarting the use or re-doing things without a considerable effort should not be possible.	*Is it possible to lose your creations or progress if you click a wrong button? Does the product support autosaving?*

As discussed in the previous sections, the aim of the learning engagement evaluation is to create a holistic view of what kind of positive experiences the product supports and identifying the things, which hinder this experience. The explanations of heuristics are formed as questions. They provide viewpoints and examples of common ways the heuristics would apply. Using questions in this manner to support the heuristics was also used by Tondello et al.(2016) in their Gameful Design heuristics. .

The framework is following the general concept of learning engagements as many faceted, multidimensional phenomena (Fredricks et al. 2012) and (following Hassenzahl 2010, 48-49) assuming that experiences can be designed purposefully to be engaging by answering psychological needs of the users. The aspects of learning engagement in this model are looking at the evaluated product from six different viewpoints.

Autonomy

The essence of the heuristics in this category is related to understanding (or creating) the goals of use and being able to make meaningful choices based on this understanding. The choices can be related to exploration of content (A1 and A5), pacing of the use (A4, A1) or being creative (A6, A1). The Gameful Design Heuristics (Tondello et al 2016) also uses a title “Autonomy and Creativity” and links *choice*, *freedom* and *self-expression* to it. The framework uses Purpose and Meaning as a separate category. In the original research by Sheldon et al. (2001) that was adopted by Hassenzahl et al (2010), Meaning and self-actualization is a separate category from Autonomy. Hassenzahl doesn't explicitly tell why he chose to leave out Meaning and self-actualization from his final six needs. However, research by Hassenzahl et al. (2013)

noticed that when people were asked to describe and rate positive experiences with technology based on what need they would fulfill, the autonomy and meaning making were often linked.

I see goals as a prerequisite for making choices in a system. The user either approves the goals set by the system or starts creating the goals themselves with the affordances of the system. In both cases, meaning-making is essential, especially in accepting the system goal: without a clear understanding of why something should be done, it is impossible to make an informed or interesting choice. Importance of goals was also present in all my reference heuristics - both HEP and Playability Heuristics for Mobile Games included a heuristic describing either clear goals or user made goals.

Competence

Feeling of being competent is inherently related to learning. Or, as Pintrich (2003) puts it: “Students who believe they are able and that they can and will do well are much more likely to be motivated in terms of effort, persistence, and behavior than students who believe they are less able and do not expect to succeed”. C4, C8 and C9 are directly related to the optimization of tasks and overcoming challenges.

Some heuristics are clearly related to usability and ease of use (C2, C3 and C7), because bad usability, such as confusing navigation, can make the user think that they are not competent enough to use the system, although the problem isn’t necessarily related to the user’s skills, but rather could be avoided with better design.

Heuristic C6 “Progression on the product depends on succeeding on things relevant for learning.” was added due to our experience in EAF that some learning products tied points or other rewards to mastering interaction mechanics, such as accurately jumping from platform to platform, and the points and progress had only a secondary connection to the topic of learning eg. correctly solving the math problems.
Several studies have also shown that gamification elements executed in this manner are mainly related to extrinsic motivation and don’t help foster intrinsic motivation (van Royn and Zaman 2017). This is also present in heuristic C1: “The product rewards the user in a meaningful way and according to the challenge”

Relatedness

In this aspect, heuristics RL4and RL6 are related to possible interaction between users. Gameful design heuristics (Tondello at al 2016) are linking multiplay to relatedness, and Korhonen and Koivisto's heuristics (2006) has a section dedicated to multiplay that informed the formulation of these heuristics. The original definition of relatedness need by Hassenzahl (2013) is also referring to interaction with other people.
The heuristics RL1, RL2, RL3 and RL5 are not referring to interaction with users but actually interaction and relationship with fictional events and characters, as well as tone of voice the user is addressed. Gameful design heuristics (ibid) categorize narrative elements to immersion, and also HEP and Korhonen and Koivisto's heuristics (ibid) have a separate category for narrative and story. The relatedness aspect comes from the connection of narrative elements to the learning situation. If user can link the narrative and events presented in the learning solution to their own experiences, they can potentially feel that the things in the solution are relatable and meaningful to them personally. However, this interpretation of relatedness would need further theoretical validation.

Respect

As mentioned before, the original term by Hassenzahl, Popularity, was changed to Respect to describe how the developers of the solution are taking the user into account (respecting them). RS1 and RS5 are related to aspects of polished and good quality product development. Heuristic for good system feedback is directly referring to Nielsen's (1994) heuristic "#1 visibility of system status", and frequent and clear feedback was also a property in all reference heuristics. RS2, RS3 and RS4 refer to taking into account users of various abilities, skill levels and backgrounds. The view on skill level is slightly different from challenge optimization under competence; Here the point is to look at things that support and scaffold users in their actions without unnecessarily getting in their way - for example, a skilled user can decide to skip a tutorial, and a user wondering what to do should always know where to get help. This is inline with Nielsen's (ibid) heuristic #7: Flexibility and efficiency of use, as well as the Lead, Follow and Get out of the way general user experience model developed in IBM (Kamper 2002). RS2 and RS3 were seen as important additions to the framework. The

learning solutions are often developed for formal education, so providing content that can be seen discriminatory or alienating a significant amount of potential users, wouldn't be considered valid in such context. However, it should be noted that what is considered discriminatory or stereotypical can greatly vary in different cultures and contexts, so considering this angle requires some expertise and interpretation from the evaluator.

Stimulation

The essence of heuristic in this category can be summarized as wanting to "not being bored". These heuristics consider visual and audio stimulation and aesthetics of the solution (ST4), but The main emphasis is put on bringing in novel content and progressive challenges. The rationale for this is flow theory, that posits that too easy challenges lead to boredom (Csikszentmihalyi 1991). In Korhonen and Koivisto's (2006) heuristics the avoidance of boredom is present in gameplay heuristics: "GP8 There are no repetitive or boring tasks" and "GP11 The game does not stagnate".

Safety

According to Hassenzahl and colleagues (2010) "Security can be understood as a ''deficiency need", i.e., a need that creates negative affect if blocked, but not necessarily strong positive feelings if fulfilled." Based on my experience with the product evaluations conducted so far, this seems to be true; Products that succeed in other categories may still fail in important safety aspects, and products that are working optimally according to this category may not produce especially meaningful positive experiences. SA1 and SA5 are directed from Korhonen and Koivisto's (2006) heuristics: "GU10 The player cannot make irreversible errors" "GP14 The player does not lose any hard-won possessions". They are also related to usability and reflect Nielsen's heuristic on error recovery. SA2 and SA4 are related to safe multiplay experience, that is also notified in Korhonen and Koivisto heuristics and HEP framework. Especially for products targeted to minors and used in schools, the issues of privacy and prevention on on-line bullying should be taken into account. SA3 refers to advertizing and "suitable content". Similar to RS3 and RS2, applying this requires knowledge of the targeted user and context of use.

4.5.1. The evaluation software

In May 2017, EAF started the development of web-based evaluation software that could be used for systematically conducting evaluations and storing the results. The first round of software development ended in October 2017, when the evaluation tool was functioning for using it with EAF employees - each user with the EAF account could log in, conduct an evaluation or edit evaluation that was filled previously, and print out the evaluation report.

In April 2018 the second development round was started. In this round, the aim was to create a separate interface for freelance teachers with more limited viewing and editing rights than for the EAF employees. The update round was finished in September 2018, and the current functions of the evaluation tool were finalized. After this round, only bug fixes and minor changes to the UI were made. In the evaluation tool, several teachers can evaluate the same product, and after submitting the evaluation, the results from individual evaluations can be compared and the admin user can select what data to export into the evaluation report that is exported to Google Slides document.

The evaluation tool works in a web browser, and contains different sections for evaluating the learning goals, pedagogical approach and learning engagement. It also has pages developed for administrative purposes, for example the Basic Information page for filling in metadata about the product. In this study I will more closely describe only the function in the Learning Engagement sections, since they are relevant for this research.

The Learning Engagement section is divided to six sub-pages according to the heuristic categories. At the top of the page, the user will see the category name and the definition for it. And underneath are the heuristics for that category. For each heuristic, there is a plus-button, where the user can view the explanation for that heuristic. The user conducting an evaluation can rate each heuristic in the scale from ‘Fully’ to ‘not at all” based on how much they agree with the statement (Fully = Fully agree and Not at all = not at all agree).

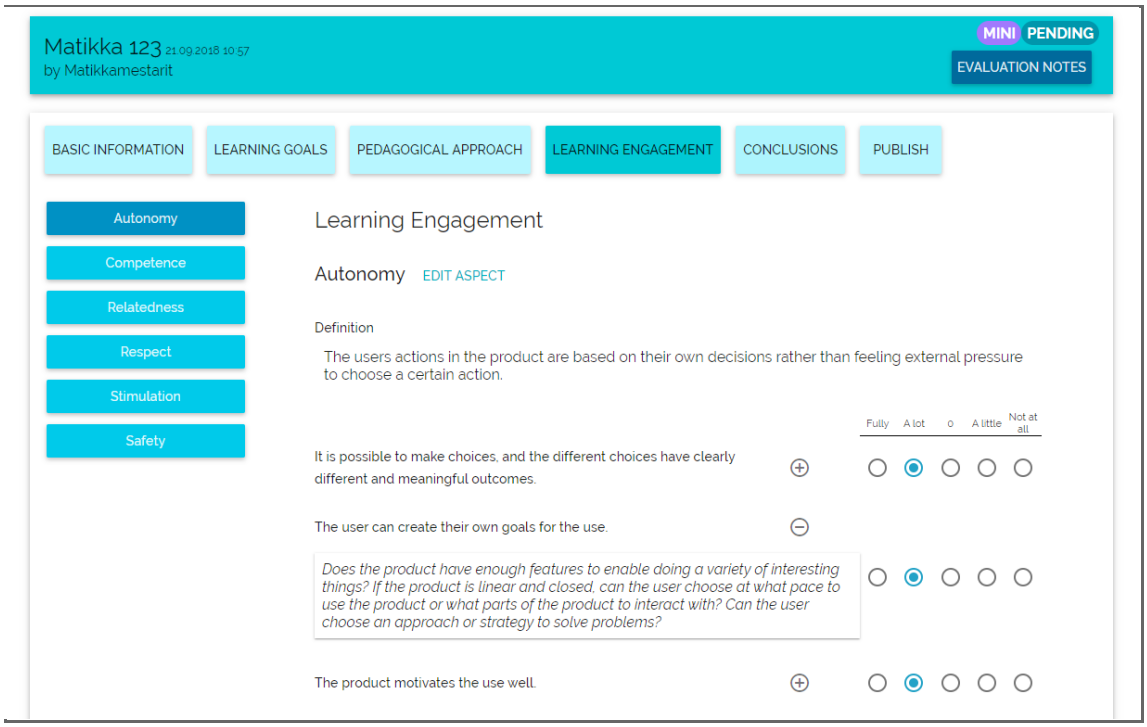


Image 2: A screenshot of the evaluation software, viewing and answering the heuristics. Education Alliance Finland, 2020.

If the evaluator finds the heuristic hard to apply for the product, they can leave the heuristic unanswered. Currently the software does not include a function for ignoring the heuristic or removing an answer - and as I will discuss later, this turned out to be a major development point.

Each Learning Engagement section also includes Strengths and Development areas text fields. The evaluators are instructed to explain their review and write down notes about the things they were paying attention to when answering the heuristic statement. It is also possible to upload explanatory images to each section. This was done to help explaining situations, where the evaluator refers to a feature that is difficult to describe, for example colors, figures, or other visual elements, or if an unexpected situation has occurred, the application has crashed, or provided error messages. For Strengths, the evaluator can also add an image of something particularly cool and inspiring that demonstrates the potential of the system.

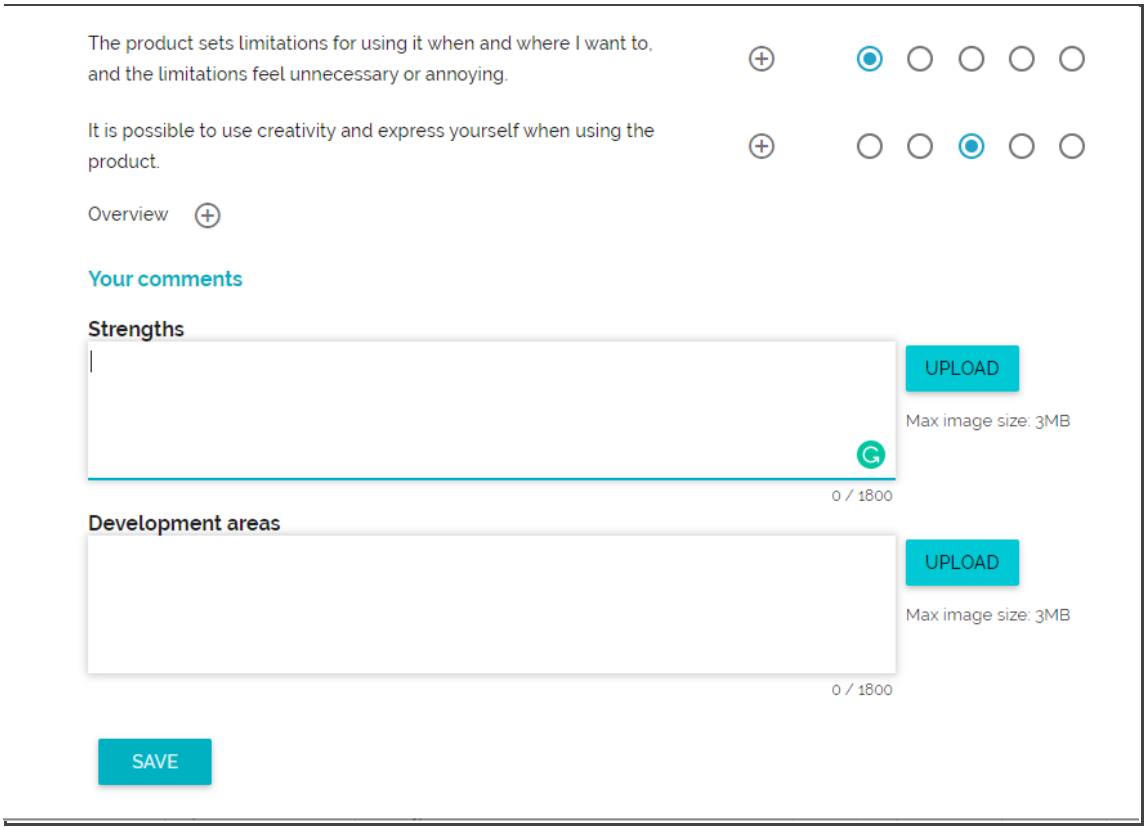


Image 3: A screenshot of the evaluation software, open text fields. Education Alliance Finland 2020.

The user's inputs are saved whenever they navigate away from the page, so they can fill in the evaluation in multiple sessions and do the sections in any order they like. Once they are ready, they can submit the whole evaluation, after which they cannot edit their answers any more.

The EAF administrator can go view the evaluation of the individual teacher. If the evaluation is submitted, the administrator can see the individual teacher's score (the answers "Fully - Not at all" are translated to scale 1-5) and compare it to the average score of all evaluators. They can also see all comments each of the teachers have submitted.

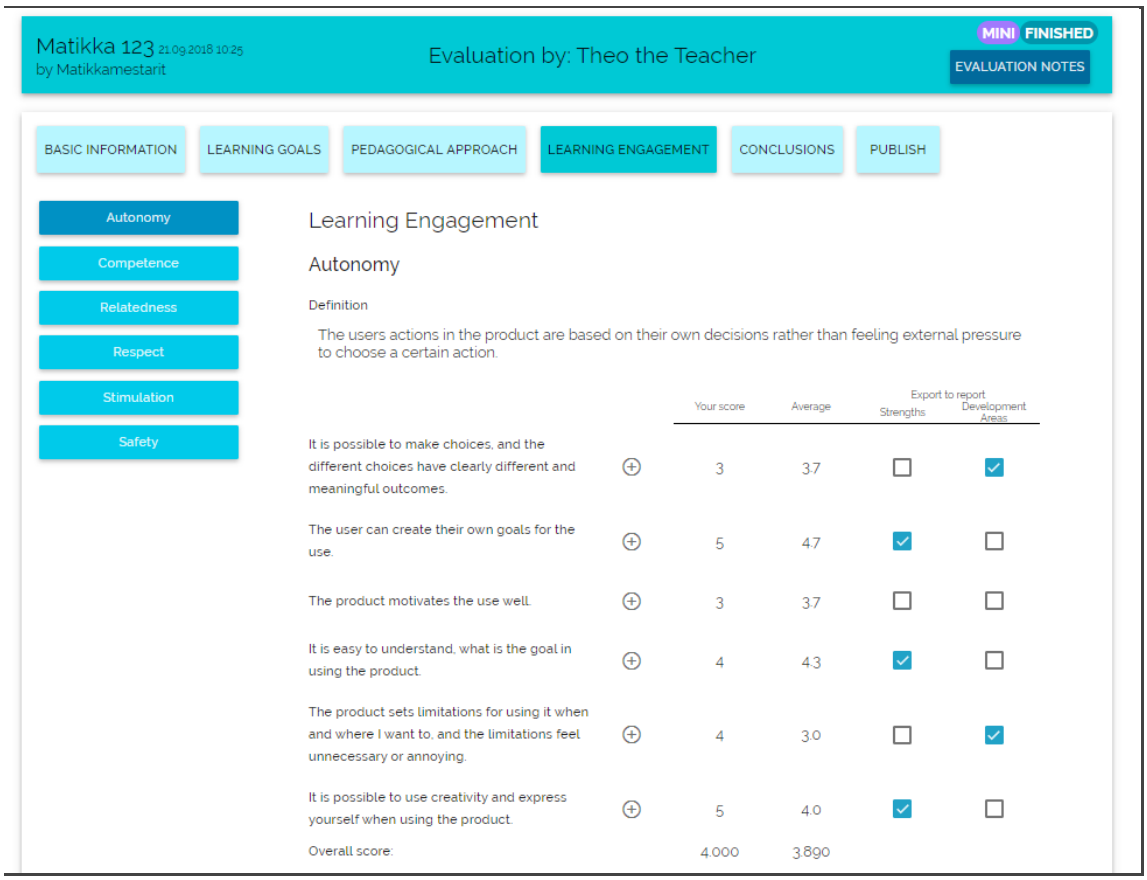


Image 4: Screenshot of the evaluation software, submitted evaluation with scores in heuristics. Education Alliance Finland, 2020.

The EAF Administrator will view all the comments and write a paragraph that summarizes the main findings, sometimes using direct quotes as well. They can also mark certain heuristics as strengths and development areas. These are then exported to a report. From each aspect, the report includes one page describing strengths (Image 5) and one for development areas (Image 6).

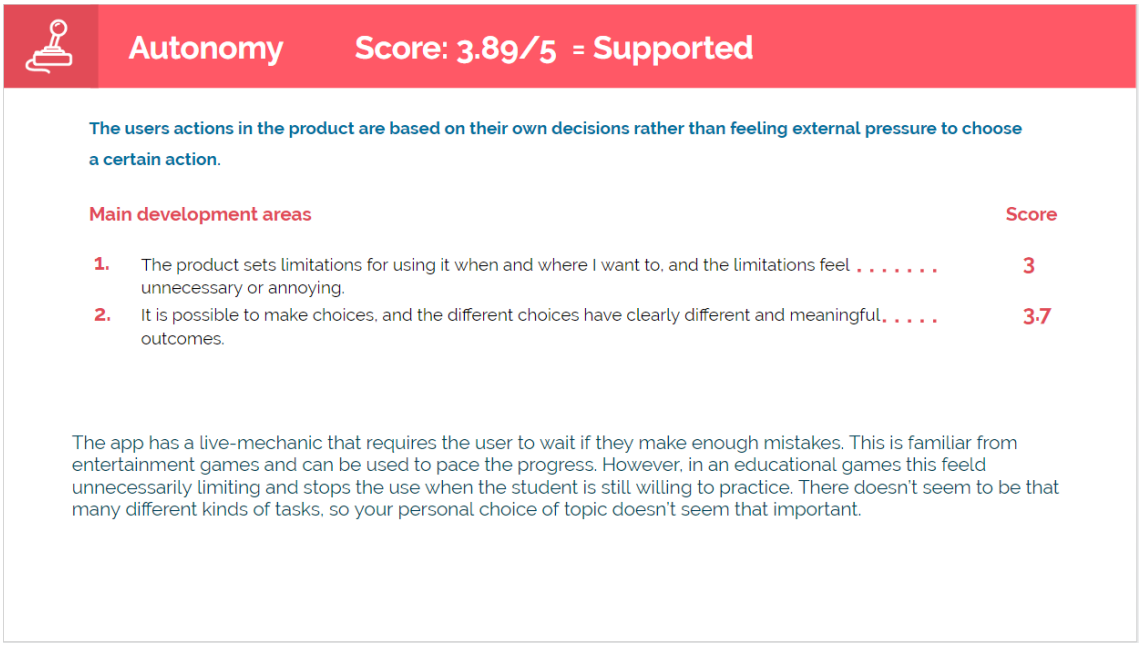
Autonomy Score: 3.89/5 = Supported

The users actions in the product are based on their own decisions rather than feeling external pressure to choose a certain action.

Main strengths		Score
1.	It is easy to understand, what is the goal in using the product.	4.3
2.	It is possible to use creativity and express yourself when using the product.	4
3.	The user can create their own goals for the use.	4.7

Matikka 123 is a product build for a clear purpose, and the learners can use it to drill certain areas of mathematics. There are plenty of open-ended tasks that require creative problem solving.

Image 5: Example of an evaluation report, Strenghts. The report page is for an imaginary product. Education Alliance Finland 2020.



Autonomy Score: 3.89/5 = Supported

The users actions in the product are based on their own decisions rather than feeling external pressure to choose a certain action.

Main development areas		Score
1.	The product sets limitations for using it when and where I want to, and the limitations feel unnecessary or annoying.	3
2.	It is possible to make choices, and the different choices have clearly different and meaningful outcomes.	3.7

The app has a live-mechanic that requires the user to wait if they make enough mistakes. This is familiar from entertainment games and can be used to pace the progress. However, in an educational games this feeld unnecessarily limiting and stops the use when the student is still willing to practice. There doesn't seem to be that many different kinds of tasks, so your personal choice of topic doesn't seem that important.

Image 6: Example of an evaluation report, Development areas. The report page is for an imaginary product. Education Alliance Finland 2020.

4.5.2. Rating scale

The rating of heuristics is not originally part of the heuristics method as developed by Nielsen. However, the evaluation process where the evaluator rates the heuristics have been used in heuristics evaluations (Albion, 1999).

After the teacher has submitted the evaluation, their rating “Fully” - “not at all” is turned to a numeric scale of 1-5. The software also calculates the average of the numeric score of all evaluators for each heuristic and for each category. The point of this is to more easily assess, in which heuristics and categories the product is performing especially well and where there is room for improvement.

An overall rating is problematic, because it gives each aspect a similar emphasis. Both in the works of Sheldon et al. (2001) and Hassenzahl et al. (2010) the experiences were categorized by the main need(s) they fulfilled, and a single experience was seen as satisfying even if it fulfilled only one possible need. Therefore it can be argued that it is irrelevant to punish a product if it fails in one or more categories but succeeds in another. Based on our practical experience with learning solutions, this could be the case with products that rely heavily on individually solving problems given by the solutions. These kinds of solutions do not often support the user's autonomy or relatedness (e.g., there are no chances for self-expression or communication with other users),but offer a good experience of competence. The evaluators always have a chance to ignore a heuristic, if they think the solution does not need the features that would be relevant for the said heuristic. However, making a judgement of when the heuristic is irrelevant is fuzzy at best and requires some expertize. This was also discovered with the research about the use of the heuristics, that is described in the chapter 5.3.3.

Another problem with overall rating is that it doesn’t reflect serious issues, if they violate only one heuristic or one category. Based on our experience with educational products, many issues, such as poor feedback on tasks, are violating heuristics in several categories and therefore lower the score in all of them - and therefore are reflected to overall score as well. Yet, there are some cases, for example, in the Safety category, where a major issue doesn’t have a major impact on score. For example, a situation where the user’s registration email address or their date of birth is shared with other users without the user’s consent is a serious privacy violation and would immediately

mean that the solution couldn't be used in schools according to regulation in several EU countries. However, this would only violate one heuristic S4: "If the user shares content - their work, their comments or anything else - it is always clear who has access to the shared content". These kinds of issues are mostly related to user's privacy or very serious usability issues that prevent the use of certain features, as well as bugs that make the usage unstable.

However, the numeric ratings of heuristics are not used in comparing products to each other, nor finding statistical relevance. The score is used to provide a *frame of reference inside the product* for the client receiving the evaluation. In this, it is similar to severity rating of usability problems found with heuristics (or other) methods. Jacob Nielsen (1994b) describes the purpose of severity rating as such:"Severity ratings can be used to allocate the most resources to fix the most serious problems and can also provide a rough estimate of the need for additional usability efforts". This is also the purpose of the rating in the Learning Engagement Evaluation.

The scale used in EAF reporting is very robust:

5-4 = **Well supported:** There are several well executed features which support this aspect of learning engagement
3-2 = **Supported:** The product takes into account this aspect of learning engagement. Some improvements could be made in order to make the support better.
1 = **Not Supported:** There are issues with the learning engagement in this area.

Based on the conducted product evaluations, products that receive generally positive comments and don't suffer from major usability issues generally score close to 3 in all categories. Products delighting the evaluators and providing exceptionally interesting and well produced content score above 4, but very rarely over 4.5. Major issues in one or more aspects can be suspected if the score drops below 2.7. However, further study on the validity and consistency of the interpretation of the scores is needed. This could be done by more rigorously analyzing the actual scores in the evaluations, but this wasn't on the scope of this research.

5 RESEARCH TO VALIDATE THE LEARNING ENGAGEMENT METHOD

As pointed out in the chapter 4.3.1., when validating a set of heuristics, the validation process is often fairly simple. Sometimes, as was the case in Gameful Experience Heuristics (Tondello et al 2016), the validation research is done only after the list of heuristics has been published (Tondello et all 2019). If the list of heuristics is validated by having UX professionals use it in an analysis and then asking their view on the topic, the amount of professionals and evaluated products can be fairly small, and the feedback they give is not analyzed with formal methods, but considered by the researchers in a more freeform manner. This was the case with all heuristic lists I used for informing my list of heuristics, and it's in line with the findings of the metastudy by Hermawati and Lawson (2016). Rusu et al. (2012) also don't have exact recommendations on how to collect, analyze and incorporate the feedback by professionals using and commenting the heuristics.

In this thesis, I have seen it necessary to study more deeply, how the users of the heuristics are using and interpreting the heuristics in the method. There are several reasons for this. First, the quality and accuracy of the evaluations conducted by EAF needs to remain consistent. Because the heuristics are used to study a variety of products, it is important to find out how well they fit each contextThe results from this, and the original validation was already established in tests carried out when the heuristics were initially formed and used by EAF team members; myself and two experts in evaluations which was done during 2016 and 2017.

Secondly, it is necessary to understand how the method and the heuristics are interpreted by evaluators who are experts in pedagogy but not necessarily in UX design or evaluation. It would be ideal to minimize the training needed for using the method. Therefore the heuristics should use unambiguous language that is targeted to teachers and not to people proficient in interaction design terminology. Thirdly, since EAF is using freelance teachers for the work, we are taking in new evaluators regularly. It is

possible that a teacher may evaluate one product that is interesting for them, and after that have a long pause before finding another one they would like to take. Therefore, the heuristics should be easy to learn initially, and even easier to apply correctly in a novel situation.

In this part of the research I am describing, what kind of issues the teachers found when using the Learning Engagement method, and this way assessing the applicability and clarity of the heuristics. The qualitative research method I am applying is thematic analysis. I am identifying the points, where the heuristics could be improved and discussing the potential requirements and competences the teacher should meet for using the framework.

5.1. How the evaluations are conducted

When a new product is given to evaluation, EAF selects a minimum of three teachers for conducting the evaluation. There is always an EAF administrator employee making sure the teachers have all the material and information they need and reviewing the product his/herself as well. The EAF administrator can be either a UX expert or a pedagogue. The selection of teachers is based on their qualifications - ideally they would be reviewing products that they can use in the level and subject they are teaching at, but often a teacher who has evaluated some products before might sign up to evaluate something outside of their regular expertize.This is not discouraged, since the evaluation doesn't necessarily require being an expert in the content the product is teaching, but of pedagogical theory and practices. The teachers should be competent in the language the product is in, although in some cases an interpreter or automatic translation is used, if it is possible to guarantee a sufficient understanding with them. The only formal qualification required is that the evaluators should hold a degree in education and have experience in teaching, even if they wouldn't be teaching professionally at the moment. Some evaluators have been accepted if they have a degree in a closely related field, such as educational technology design, and some practical experience in teaching. Also, some students close to graduation have performed evaluations.

When the evaluators are selected, EAF delivers the necessary materials to them. This can be credentials to the evaluated system, software to download, digital documents, or a physical product that will be posted. In cases where the system is very complex or broad, the teachers can have an introduction to the system/product from the company of origin or EAF. Expert evaluations, including heuristic methods, do not require the evaluator to start using the solutions without help, and helps and hints should be available also during the process, if the evaluator has questions (Nielsen 1994c). In EAF evaluations the evaluator should use the necessary means to find out how the product works and when assessing the first-time experience and learnability, make judgements based on their expertise rather than documenting their actual experience.

Usually EAF reserves 2-3 weeks for the teacher to conduct the evaluation. They can use the product independently or, if desired, use it with their students or colleagues. Sometimes the usage period can be extended, but as a rule the aim is not to conduct a study on long term impacts of the use. It is possible that a teacher may be familiar with the product before doing the evaluation and in that way has gained an expertise on the usage. This may have an impact on their judgement, but whether it gears the results to a more positive or negative direction can't be said for sure but would be an important point to analyze further. An expert user is likely to be more familiar with all features and use cases, but it is necessary to make sure all novice users will also form a realistic, holistic view of the product before conducting the evaluation.

The evaluation is done in the EAF browser-based evaluation software. The features in Learning Eengagement Evaluation were described in the chapter 4.5.1. After finishing the evaluation, the EAF administrator, who has also assessed the product, will go through the individual evaluations - the teacher's scores, individual answers and textual comments. If he/she finds outliers in the ratings or a comment that requires further explanation, they will contact the teacher to ask for clarification. In some cases, the teacher and the EAF administrator may agree for some answers to be changed. This can happen for example if a teacher has missed a feature or if they have misinterpreted the question. The teacher can be also asked to iterate their evaluation.

When all evaluations are approved by the EAF administrator, he/she will form a report of the findings. The report is gone through with the client, and sometimes the feedback

from the client is shared with the teachers for training purposes and sometimes to clarify issues the teachers may have encountered during the evaluation.

5.2. The method of analysis and data

Thematic analysis is a foundational method of qualitative research. It is a method for identifying, analysing and reporting patterns (themes) within data, and it can be utilized from various theoretical and epistemological standpoints. It can be conducted inductively, when the analyzed data is coded without existing theoretical coding frame, or theoretically by tying the coding process to a certain theory. (Braun and Clarke 2006)

When conducting thematic analysis, it is important to justify and openly present the decisions made in each step of data collection and interpretation, because there aren't rigid steps proposed by the method (Guest, MacQueen and Namey 2012, chapter 10). Thematic analysis can be applied to a variety of data types; interview transcripts, texts, media broadcasts and so on (Braun and Clarke 2006). In thematic analysis there are no clear guidelines on how many participants or researched data points the research should have for validity. As thematic analysis is a flexible qualitative method, and the aims of the research define, when the data salient and is sufficient for reliable interpretations (Guest, MacQueen and Namey 2012, chapter 3).

In this study, the main limitation for the collection of data was time. The collection happened between January 2020 and April 2020, so the selected evaluations were the ones worked on during that time. Additional data could have been taken in by analyzing existing evaluations. However, with them, it would not have been possible to discuss with the teachers about their findings and interpretations to confirm potential problems, or at least the conversations would have been unreliable because some time had passed after the evaluation. The collected data was also deemed sufficient, because problems with the same heuristics started to repeat and based on my previous experience, the problems represented in the data covered all issues that I had suspected so far, and some more.

5.2.1. Collecting the data

The data I am using in my analysis is about the problems the teachers encounter when using the heuristics. The Learning Engagement Evaluation framework has been used by Finnish freelance teachers since September 2018, and during the use some problems were discovered, but they weren't systematically documented. However, each evaluation was validated by an EAF expert and the teachers were trained in the work to use the heuristics correctly. The notified problems were also taken into account in training of new teachers. The systematic documentation of problems for this study started in January 2020 and was finished in April 2020. At that time, EAF was conducting a pilot project in Switzerland, and trained Swiss teachers to create the evaluation. Therefore it was possible to collect experiences of several novice evaluators starting to use the framework.

The process of collection happened in a following way:

- The teacher creates an evaluation in EAF evaluation tool. They rate each heuristics based on their view of the product, or may choose to ignore the heuristic. They also wrote open feedback about the product.
- I examined the teacher's answers and based on my knowledge of the product looked for inconsistencies in their answers that could imply a problematic heuristic. This was done by 1) comparing the answers of 3 teachers evaluating the same product 2) comparing the rating of the heuristics to the text comments of each teacher. When discussing the results, I will describe more closely how each type of a problem was identified.
- I contacted the teacher either via email, phone call or instant messaging service such as Whatsapp, Messenger or Skype chat and asked how they interpreted the heuristic and what was their reasoning behind their rating or comment.
- To form a data point, I marked what the heuristic was, why it caused me to suspect a problem and what was the teacher's rationale behind their answer.

The collection and storing of data was done in a Google sheet as text. The original emails or instant messages with teachers were not stored for research purposes, but direct quotes from them were included to the sheet.

The evaluation was conducted to 14 products from the following categories; Gamified teaching tools, digital solutions for learning some content (spelling and grammar, mathematics, humanities and economy and entrepreneurship), a programmable robot and related teaching material and programming environment, content creation tools and related learning material (programming, design, 3D modelling) The evaluators consisted of 9 Swiss teachers, 3 Danish teachers, 11 Finnish teachers. The Swiss and Danish teachers were new to the evaluation, 1 Finnish teacher was a first timer, 10 had done at least one evaluation before, some being experts with more than 10 evaluations done.
Each of the products was evaluated by three teachers, resulting to 42 individual evaluations. Majority of evaluations were found problem free. In 19 evaluations it was necessary to suspect problems related to the use of heuristics, resulting to 44 individual suspected problem situations that were then confirmed with the teacher. In the end 40 of these were actual problems in the use of heuristics.

5.2.2. Coding the data and finding the themes

The book Applied Thematic Analysis (Guest, MacQueen and Namey, 2012, chapter 2) describes, that the procedure of code creation can happen iteratively; first reading the text, writing notes and proposing possible themes. Then, refining the themes to codes with clear definitions. The validity and non-ambiguity of the codes should be verified by having several people coding the same text with the codes, and if there are differences, the codes should be iterated. In case there is only one researcher, it would be possible to test the codes also by doing the coding twice with a week or more in between.

In this study, I created the codes and evaluated their usefulness myself. When collecting the data, I had already done plenty of interpretation. The actual raw data I obtained are the teacher's comments, emails, and instant messaging conversations. However, without the context - which specific product and heuristic were we talking about - these would be fairly useless material analyses. To form my final data set I ended up summarizing each problem to a short note that describes the situation and my conversation with the

teacher, sometimes using direct quotes. In some evaluations there is a recurring problem that is very easy to identify. For example, a teacher consistently interprets heuristics written as a negative sentence (including a word “not”) the wrong way around - they rated the heuristic “not at all”, even though they agreed with it. If this happened several times with the same teacher in different heuristics or different evaluations, I have confirmed the error once or twice, but marked it as a separate problem also for similar heuristics which I haven’t specifically discussed with the teacher. The final data was an excel sheet that was exported to pdf and analyzed in Atlas.ti qualitative analysis software (vs.8.0).

Because the data was already formulated in a way that each problem was shortly summarized, the natural way to approach coding was to assign at least one specific code to each problem. Because the amount of problem descriptions was still rather easy to manage (44 individual descriptions, resulting to 4 pages of text), I started by examining each of the problem descriptions and writing a code that described the root reason for it. These were for example, “misunderstanding a term”, “No relevant features” or “lack of examples”. After the first round of coding, I iteratively went through the material again, and after being satisfied with the codes, started grouping them in Atlas.ti. Grouping resulted in final identification of themes that I will describe in the findings.

5.3. Findings and discussion

In the end, if was possible to identify six categories of themes from the data. Five of them describe actual problem situations, and one is a situation where a problem was suspected, but it wasn’t actually a problem in understanding or interpreting the heuristic but simply a difference of opinion that standed out. Next I am presenting all six categories.

5.3.1. Heuristic written as a negative sentence

By far the most common problem in applying the heuristics happened when rating heuristics that were written as negative sentences (including the word “not”). There were 12 separate occasions where the teacher had answered “not at all”, although they agreed with the statement. Most often this happened with the heuristics in Respect category: “The product doesn’t make assumptions on the player's age, gender, race or

origin.” and “The product doesn’t include discriminative narrative or enforce unnecessary stereotypes”. This problem was often easy to identify, because with these statements the teachers tended to mark their grading to the end of the scale (“not at all”), and very rarely the products had content that could be considered stereotypical or discriminative. Often the teachers also commented on the fact. For example, an teacher who had marked all negative sentence heuristics as “not at all”, commented: *“[The product] runs very stable and reliable. I have not found any discriminations (sic).”*

Very often a teacher had systematically answered all of the negative sentence heuristics “the wrong way around” , so it was clear to suspect a problem. For some reason “The product doesn’t have bugs which cause errors or crashing. “ was more often graded correctly, even by an evaluator who mis-graded the two other Respect heuristics written as negative sentences. The reason for this was not clear, but it could be suspected that a shorter sentence was easier to understand correctly. Also the Safety heuristic: “The product doesn't include content or advertising which would be harmful for the targeted users” was incorrectly interpreted less frequently.

5.3.2. A Better explanation of the heuristic is required

Under this theme, I counted situations where the problem occurred for two main reasons: Lack of explanation or example application of the heuristics or heuristic using terminology that was not unequivocal to the teachers.

Most of the teachers were novice evaluators, so they encountered and interpreted the heuristics for the first time in the context of the product that they were evaluating. In some cases, this led to situations where the teachers interpreted the heuristic or some word in it in a manner that wasn’t completely correct, because they lacked explanation or examples of broader context.

In the Competence heuristic “The product’s challenge level is optimal for the targeted users, or it can be chosen” was problematic. ‘Targeted users’ was interpreted very broadly, or sometimes even incorrectly. Some teachers did not consider the target group defined by the product developer and judged this, for example, based on the level they

teach themselves. However, on some occasions the target group was not clear, so the teachers needed to do some interpretation.

“Challenge” was also interpreted in various ways. In some evaluated products the challenge set to the students greatly depended on the teachers - for example, a product for creating gamified quizzes where the teacher could either make a quiz themselves or use and modify some of the ready-made quizzes - so the challenge level could be chosen. However, a teacher commented on the challenge level of ready-made materials and based her evaluation on that.

Another heuristic that was related to target group was under Respect: “The product doesn’t make assumptions on player’s age, gender, race or origin.” In great majority of cases this was interpreted correctly (although sometimes the grading scale was understood the wrong way around). In one case, the teacher marked this as “0”, because he thought the app is targeted to small kids, so it is setting an assumption. This heuristic is only looking at situations, where there's assumptions that are harmful, unnecessary or too limiting, it should not be punished for having a clear target group

In another case, a teacher had commented that the product is too discouraging for weak learners, so it *“[makes assumptions on] learning abilities”.* However, in this case the teacher had still marked the heuristic as “fully agree”, which implies she understood the correct application of the heuristic, but still wanted to comment on a topic that the heuristic didn’t take into account.

A term that was often understood incorrectly was “feedback”. In Respect, the heuristic “The product gives clear feedback on all your actions” was often understood to consider the pedagogical feedback, such as ways how the system tells the learner, why their solution to problems is correct or wrong, and more technical system feedback (for example, notifying the user about saving of content) was overlooked. This interpretation became apparent in evaluations of products that had very little interactive features for learners - for example, they included resources for teachers or the content wasn’t really interactive at all. In these cases, some teachers marked this heuristic as “not at all” or “little” because they thought the system doesn’t give pedagogical feedback. When explained that heuristic referred to all feedback, including system messages and feedback on teachers’ actions, the evaluators corrected their judgements. This finding

implies that in some evaluations the teachers might not have been reporting problems related to system feedback, if the pedagogical feedback has been good. Verifying this possibility would require analyzing more evaluations and asking this from the teachers.

5.3.3. The Heuristic is difficult to apply, because the product lacks some features

In this category, the problems for rating the heuristic happened because the product didn't have any features related to the situation the heuristic refers to. This happened especially with the Safety heuristics related to social features: "There is a way to report and possibly block misbehaving users" and "If the user shares content - their work, their comments or anything else - it is always clear, who has access to the shared content." If the product didn't have features related to sharing or communicating with other users online, these heuristics would have been reasonable to ignore. Some teachers didn't either remember that it was allowed to ignore a heuristic, so they gave a rating (usually "0") or they accidentally marked an answer that they couldn't remove any more. This is recognized as a major flaw in the software. It would have been good to include a clear "not applicable" or "ignore the heuristic" option to avoid this kind of situation. This type of problem was easy to identify, because it was clear, when the product didn't have relevant features.

More ambiguous problem situations occurred in cases where the teacher would have wanted to have features in the evaluated solution that the heuristic referred to. For example, in Relatedness there is a heuristic "The product supports communication with other people and there are good reasons to communicate" and "The product supports social interaction, such as multiplay or sharing of content with other people". In one evaluation, two teachers marked these heuristics as "not at all" or "a little", resulting in a score of 1 or 2. The product in question was an interactive story book for preschool kids, so it had a very narrow use case. Asking for social communication to be part of the product seemed a bit far fetched, since these features would take a lot of time to develop and they would alter the use case of the product significantly. Therefore it is questionable whether the product should be "punished" for lacking these kinds of features. There were also cases where this type of interpretation was justified, for example for products targeted to classroom use or remote learning in school context. In

those products, adding (online) social features would be a significant improvement, and marking these heuristics as problem areas was justified. Thus, applying a heuristic that refers to features that the product doesn't currently have, requires some expertise from the teacher; they should have a view on how much the nature of the solution changes if the features are developed.

5.3.4. A problem was identified incorrectly

In a small number of cases where I suspected a problem in interpretation of a heuristic, the teacher had identified a problem, but reported it in a relation to a heuristic that didn't really fit this problem.

For example, when rating a heuristic "The product's challenge level is optimal for the targeted users, or it can be chosen" a teachers referred to the quality of challenge: *"[T]he product is directed to a very specific group with high mathematical and abstract thinking skills. Other kids [who are more oriented to arts or humanities] do not find too much challenges in [the product]".* The heuristic doesn't really take a stance on whether the challenges are interesting or varied - rather, if the challenges are achievable and within the skill level of the targeted users. The quality and topics of the challenges in the product can be assessed in relation to for example Autonomy and Stimulation heuristics ("The product motivates the use well." "The product encourages exploring it further. ").

These kinds of application problems were rare and they could be identified from free comments left by the teacher. However, without more extensively questioning the teachers about their rationale when rating each heuristic, it is not possible to say for sure, what exact properties of the product the teachers were thinking in their ratings. Therefore it is possible that some ratings were based on judging the wrong properties of the product, but this was not possible to identify because the teacher didn't describe their thinking extensively enough in the comment. All in all, this was a problem in some of the evaluations; Some teachers put a lot of effort in explaining their thinking, but some left only very short and generic textual comments.

5.3.5. A relevant issue had been missed

When considering usability heuristics, missing a relevant problem is a bigger issue than reporting a problem that might not be a relevant problem. Therefore it is necessary to have more than one person to inspect a product. In EAF evaluation the aim is not to find and report all possible usability problems, but find bigger themes and factors about the user experience. However, it is still possible that some major issues or relevant problems can be left unnoticed, at least by part of the evaluations.

Within the analyzed evaluations, there were two cases, where some of the teachers had missed a relevant, problematic feature in the product. The evaluated system had plenty of ready made content and one teacher had run into content that they considered to be stereotypical and not in line with her understanding of what should be taught, and notified this in their rating of heuristic: “The product doesn’t include discriminative narrative or enforce unnecessary stereotypes”. However, not all teachers viewed this content, so they didn't note that in their rating. When confirmed, they admitted they would have agreed that the content is stereotypical

In the second case, all three teachers ignored a feature related to account creation in the product. When the user creates an account in the product, they set an account name. However, the name cannot be changed later and it is shown publicly. This violated the Safety heuristics:”If the user shares content - their work, their comments or anything else - it is always clear, who has access to the shared content. “ “The user cannot make irreversible errors. Points that lead to restarting the use or re-doing things without a considerable effort should not be possible”. The issue was found by me when administering the evaluation. This case demonstrated that a teacher who is not an expert in usability inspections might not test the product as extensively or consider the same issues as relevant. All teachers admitted they had noticed that the user name was shown publicly, but didn’t think to check if they could change that and didn’t see it as an issue.

In one case, all evaluators (including myself) had missed a feature that allowed a teacher to give more additional assignments to their students. This would have affected ratings of several heuristics. The reason for missing the feature was related to both

tutoring and usability - the feature was behind the settings menu, and was communicated in a way that made it slightly unclear to understand.

Problems related to missing content were possible to identify from comments and from ratings that differed between the three teachers. In the case of the account name, all teachers had marked the said heuristics as “fully agree,” which made me confirm if they had noticed the issue.

5.3.6. Difference of opinion or style of grading

Sometimes a suspected problem in the use of a heuristic was not a problem after all. On these occasions, an individual teacher had rated the heuristic differently (more or less critically) than how the other teachers or myself saw the situation, which therefore led to me suspecting a misunderstanding or misuse of the heuristics. When asking about the teacher’s reasoning, they provided a better explanation of their thinking. This then confirmed that they had interpreted the heuristic correctly but simply had a different view about the severity of the problem. Presumably, the heuristics are robust enough to have few misinterpreations. In each product, there is some variation in the teachers’ answers. This was expected, and very rarely (in 4 cases) led to wrongly suspecting a misinterpretation of the heuristics.

5.3.7. Conclusions and ways of overcoming the problems

When analyzing the evaluations of the teachers, the overall findings were quite positive. The number of problems was quite low, and the great majority of them were repeating problems with the heuristics written as negative sentences. Many of the problems can be also overcome without altering the heuristics themselves.

The major need for improvement is in the training of the teachers; the training material should include more examples of the correct application of the heuristics, and these should be discussed with the teachers in face-to-face introductions to the method. An even more important factor is discussing the evaluations with the teachers after they have finished them - giving feedback and asking for additional explanations will also improve the teachers’ understanding of the correct usage of the heuristic. During the

period of data collection for this study this was done more systematically and extensively than before it, and many of the feedback practices should be continued in the future.

To help with the identification of the misunderstandings the teachers should be required, and encouraged, to leave more textual feedback and referring to the actual features of the product in it. Based on the experience since the start of the work with the teachers in 2018, this aspect will improve as the teachers' general understanding and experience of conducting the evaluations grows: teachers who have conducted more evaluations put more effort into writing the text. They are either more enthusiastic about the work in general or because they are more familiar with the heuristics, they can better focus on their textual feedback.

When introducing the product to the teachers, it should be made sure all evaluators are familiar with all features and content. This should be also checked when verifying the evaluations by discussing with the teachers. As a final verification, the results are discussed with the client, and in the discussion it should be made sure all relevant parts are taken into account.

To prevent the problem where the teachers are rating irrelevant heuristics or heuristics they are not really sure about, an ignore option should be developed to the software. Before this is done, it should be emphasized in the training that the heuristic can be ignored by leaving it empty.

The only issue, where alteration of heuristics could be considered, is related to the negative sentences. These have been problematic ever since the evaluation software was first developed in 2017. When doing the rating, the evaluator will consider not only how much they agree with the heuristic (fully - not at all), but also, what does it mean when they eg. agree fully - ie. how do they interpret the rating. In most of the heuristics, when an evaluators fully agrees with the statement, it means that they view the product in a positive light (eg. "The product gives clear feedback on all your actions" - "Fully agree" implies that feedback is good). The same is true also if the statement is a negative sentence ("The product doesn't have bugs which cause errors or crashing." - "Fully agree" means that product works flawlessly). This should make it easier to rate the

heuristics, because the evaluator doesn't need to consider which way to interpret the scale in each statement.

However, the method has an exception to this rule: "The product sets limitations for using it when and where I want to, and the limitations feel unnecessary or annoying." in Autonomy. Formulating this sentence as negative would have made it fairly complex. Based on the data in this study, this heuristic didn't cause any problems - the evaluators all answered to it in a logical manner. This would support changing the negative sentences to positive form. However, changing this in the evaluation software would affect the ratings of old evaluations. Therefore it should be carefully considered, especially since this problem mostly happens with novice evaluators when they first encounter the method.

6 CONCLUSIONS

This thesis both documented the process of creating the Learning Engagement evaluation method and presented a research for additional validation. The creation of literature review on the topic describes the theoretical foundation of the method, and the thematic analysis of problems the teachers encountered while conducting the evaluations provided further validation and insight to points of improvements.

The research question, how to use the need fulfilment as a theory for heuristic evaluation of user experience, was thoroughly answered. Need fulfilment has been applied to analysis of user experiences both in UX research (Hassenzahl et al 2010, 2013; Ford, Wyeth and Johnson 2012) and in games research (see meta-analysis by Mora et al 2017, 537) , and it has been conceptualized through motivational affordance (Zhang 2000, Deterding 2011). Also heuristic frameworks based on need fulfillment have been developed previously (Tondello et al 2016, van Roy and Zaman 2017). The needs presented in Self-determination theory (Ryan and Deci 2000) were often applied in these frameworks, although also various other needs have been suggested (Hassenzahl et al. 2010).

As discussed, experience is always contextual, situational and subjective (Hassenzahl 2010, 6, 10-11). Therefore any method that is based on findings of experts instead of observation of actual users can produce reliable results only within limitations. Expert evaluations still have value, since they are relatively fast and cost effective to conduct. Experts can also formulate suggestions and critique differently than users, and because they are paid to conduct their work, and using a formal method, it can be expected that their comments are more thoroughly thought of than those of users. The teachers doing the evaluation are selected based on their profession. They should know the context of use well and be able to give feedback as experts of the domain. In the use of educational technology, the situation in different countries and even schools can be drastically different - infrastructure, culture, skills of the teachers and requirements and regulations of national governments vary (European Commission 2019a). Therefore it is important to present these limitations of knowledge also to the client. And of course, it's not possible to completely know the context of an individual learner.

The study of 42 evaluations provided good insight on the second research question; How the teachers creating learning engagement evaluations are using the heuristics and what are the possible points of improvements. The conclusions of the study is that some concepts need more explaining and preferable examples of applying them, especially when the terminology is more UX or games related, since the teachers tended to view them from the perspective of their own profession. However, in general the teachers managed to apply the heuristics well, and the problems were easy to identify. It should be noted that a majority of the problems were found from evaluations conducted by complete novice evaluators. When the teacher had performed at least one evaluation before and received feedback from it, the amount of misunderstandings dropped. Therefore more attention should be paid to training and instructing.

The data that is provided in each evaluation would have chances for plenty of additional research. The open textual comments could be used to analyze several things: what features the teachers are paying attention to in the solutions? How do they perceive gamification in learning, do they comment gamified features in positive, negative or neutral light? How collaboration, competition or self-expression are viewed? This type of qualitative studies would give good insight on the use of educational technology and it’s perceived value among teachers. As the amount of evaluations grows, the data would become more rigorous over time.

As a future research topic it would be interesting to find out how well the method applies to different kinds of learning solutions, if there's differences and are some important areas of experience ignored by them. Based on the experience in all product evaluations and the smaller set of analyzed evaluations, it is possible to already give an intuitive answer; The method provides holistic feedback and is suitable for products that fit the criteria: Interactive solutions aiming to teach either a skill or a knowledge, that is used by learners (closer description of the criteria on chapter 4.1.). Analyzing this further could be done through collecting more client feedback or conducting comparative evaluations with another set of heuristics or by comparing the heuristics to results from user testing. This would also provide further validation of the heuristics, and potentially bring out things to improve in the listing. Conducting this kind of research would be also in the interest of EAF, given the resources.

Since the Learning Engagement Evaluation method is presented openly in this research, other researchers and UX practitioners could apply it in their work. Applying it to products other than learning solutions should be possible, since only few of the heuristics is referring to learning and they are not formulated to fit only eg. formal educational setting.

A major question and topic of future research would be, if and in what sense pedagogical quality correlates with engagement factors. As presented in the literature view, engagement has correlation to academic success. Yet, this has been studied mostly in the context of formal education, not as much in relation to use of edtech. The EAF evaluation method that combines evaluation of pedagogical quality and assessment of engagement is somewhat unique in the field. Comparing the results (both numeric scores and qualitative comments) from both methods could then provide evidence to this question.

REFERENCES

Alavi, Maryam and Dorothy E. Leidner. 2001. "Research Commentary: Technology-Mediated Learning - A Call for Greater Depth and Breadth of Research." *Information Systems Research* 12 (1) (03): 1-10. https://libproxy.tuni.fi/login?url=https://search-proquest-com.libproxy.tuni.fi/docview/208164214?accountid=14242.

Albion, Peter R. 1999. "Heuristic evaluation of educational multimedia: From theory to practice." In *Proceedings of the 16th Annual conference of the Australasian Society for Computers in Learning in Tertiary Education*: 9-15 http://www.ascilite.org.au/conferences/brisbane99/papers/albion.pdf

Appleton, James J., Sandra L. Christenson and Michael. J. Furlong. 2008. "Student engagement with school: Critical conceptual and methodological issues of the construct". *Psychology in the Schools*, 45: 369–386.

Arhippainen, Leena. 2013. "A Tutorial of Ten User Experience Heuristics." In Proceedings of International Conference on Making Sense of Converging Media (AcademicMindTrek '13). Association for Computing Machinery, New York, NY, USA:336–337. https://doi-org.libproxy.tuni.fi/10.1145/2523429.2523491

Ashwin, Paul and Debbie McVitty. 2015. "The Meanings of Student Engagement: Implications for Policies and Practices." In *The European Higher Education Area*, edited by Adrian Curaj, Liviu Matei, Remus Pricopie, Jamil Salm and Peter Scott. Springer, Cham https://doi.org/10.1007/978-3-319-20877-0

Bozkurt, Arhan and Gurhan Durak.2018. "A systematic review of gamification research: In pursuit of homo ludens." *International Journal of Game-Based Learning (IJGBL)*, 8(3): 15-33.

Braun, Virginia and Clarke, Victoria. 2006. “Using Thematic Analysis in Psychology.” *Qualitative Research in Psychology* 3 (2): 77–101. doi:10.1191/1478088706qp063oa.

Burke, Lisa A. and Karen E. James. 2008. “PowerPoint-Based Lectures in Business Education: An Empirical Investigation of Student-Perceived Novelty and Effectiveness.” *Business Communication Quarterly* 71, no. 3 (September 2008): 277–96. doi:10.1177/1080569908317151

Chauhan, Sumedha. 2017. “A meta-analysis of the impact of technology on learning effectiveness of elementary students.” *Computers & Education*, Vol. 105: 14-30. https://doi.org/10.1016/j.compedu.2016.11.005.

Chen, Jenova. 2007. Flow in games (and everything else). *Communications of the ACM*. Vol 50 no.4: 31-34.

Christenson, Sandra. L., Amy L. Reschly. L., James J. Appleton, S. Berman, D. Spanjers and P. Varro. 2008. “Best practices in fostering student engagement.” In, *Best practices in school psychology* (5th ed.). Edited by A. Thomas & J. Grimes Bethesda, MD: National Association of School Psychologists.

Clark, Richard E. 1994. “Media Will Never Influence Learning.” *ElR&D*, Vol. 42, No. 2: 21-29.

Cockton, Gilbert and Alan Woolrych. 2001. “Understanding inspection methods: lessons from an assessment of heuristic evaluation.” In *Proceedings of people and computers XV: joint proceedings of HCI 2001 and IHM 2001*. Berlin: Springer-Verlag: 171–192.

Crookall, David. 2000 “Editorial: Thirty Years of Interdisciplinarity.” *Simulation & Gaming* 31, no. 1 (March 2000): 5–21. doi:10.1177/104687810003100101.

Csikszentmihalyi, Mihail. 1991. *Flow: The Pyschology of Optimal Experience*. New York:Harper Collins Publishers

Digital Education Action Plan. Available at: https://ec.europa.eu/education/education-in-the-eu/digital-education-action-plan_en
Accesses 8.3.2020

Deci, Edward L., Richard M Koestner and Richard Ryan. 1999 "A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation". P*sychological bulletin* 125, 6 (1999): 627–668.

Deci, Edward L., and Richard Ryan. 2002. "Overview of self-determination theory: An organismic dialectical perspective." In *Handbook of self-determination research*, edited by Edward L Deci and Richard Ryan. Rochester, Ny:University Rochester Press: 3-33

Deterding, Sebastian. 2011. "Situated motivational affordances of game elements : A conceptual model." *CHI Gamification Workshop* 2011: 3-6.

Deterding, Sebastian. 2017. "The pyrrhic victory of game studies: Assessing the past, present, and future of interdisciplinary game research." *Games and Culture* 12 (6): 521–543. DOI:10.1177/1555412016665067

Desurvire, Heather, Martin Caplan, and Jozsef A. Toth. 2004. "Using heuristics to evaluate the playability of games." In CHI '04 *Extended Abstracts on Human Factors in Computing Systems (CHI EA '04)*. Association for Computing Machinery, New York, NY, USA: 1509–1512. DOI:https://doi.org/10.1145/985921.986102

Esteban-Millat, Irene, Francisco J Martínez-López, Rubén Huertas-García, Antoni Meseguer, Inma Rodríguez-Ardura. 2014. "Modelling students' flow experiences in an online learning environment." *Computers and Education* 71: 111–123. DOI:https://doi.org/10.1016/j.compedu.2013.09.012

European Commission. 2019a. "2nd Survey of Schools: ICT in Education – Objective 1: Benchmark progress in ICT in schools." Luxembourg: European Commission. doi: 10.2759/23401
Available at https://ec.europa.eu/digital-single-market/en/news/2nd-survey-schools-ict-education
Accessed 8.3.2020

European Commission. 2019b. “2nd Survey of Schools: ICT in Education – Objective 2: Model for a ‘highly equipped and connected classroom’”. Luxembourg: European Commission.doi: 10.2759/831325 Available at https://ec.europa.eu/digital-single-market/en/news/2nd-survey-schools-ict-education
Accessed 27.4.2020

Ford, Matthew, Peta Wyeth and Daniel Johnson. 2012. “Self-determination theory as applied to the design of a software learning system using whole-body controls.” In *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI '12)*. Association for Computing Machinery, New York, NY, USA, 146–149. DOI:https://doi-org.libproxy.tuni.fi/10.1145/2414536.2414562

Fredericks, Jennifer A., Phyllis C. Blumenfeld and Alison H. Paris. 2004. “School engagement: potential of the concept, state of the evidence”. *Review of Educational Research,* 74, 59-109.

Fredricks Jennifer A. and Wendy McColskey. 2012. “The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments.” In *Handbook of Research on Student Engagement*. Edited by Sandra L. Christenson, Amy L. Reschly and Cathy Wylie. Springer, Boston, MA

Guest, Greg, Kathleen M. MacQueen, and Emily E. Namey. 2012. *Applied Thematic Analysis*. Thousand Oaks, CA: SAGE Publications, doi: 10.4135/9781483384436.

Grudin, Jonathan: “A Moving Target: The Evolution of Human– Computer Interaction”. In *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications, Third Edition*, edited by Julie A. Jacko, CRC Press LLC, 2012. ProQuest Ebook Central: xxvii-lxi http://ebookcentral.proquest.com/lib/tampere/detail.action?docID=911990.
Created from tampere on 2020-04-25 14:38:16.

Hassenzahl, Marc, Sarah Diefenbach and Anja Göritz. 2010. “Needs, affect, and interactive products–Facets of user experience.” *Interacting with computers*, 22(5): 353-362.

Hassenzahl, Marc. 2010. *Experience design: Technology for all the right reasons*. In series *Synthesis lectures on human-centered informatics, #8* Edited by John M. Carroll. Morgan & Claypool

Hassenzahl, Marc, Kai Eckoldt, Sarah Diefenbach, Matthias Laschke, Eva Lenz and. Joonhwan Kim. 2013. "Designing moments of meaning and pleasure. Experience design and happiness." *International Journal of Design*, 7(3).

Hermawati, Setia and Lawson, Glyn.2016. "Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus?" *Applied Ergonomics*, Volume 56, 2016: 34-51, https://doi.org/10.1016/j.apergo.2015.11.016.

Hietajärvi, Lauri V.0., Maksniemi, Erika. 2017. "How to Design Engaging Educational Solutions?" ELE Finland Oy / Kokoa Standard. Available at: https://educationalliancefinland.com/sites/default/files/how-to-design-engaging-pedagogical-solutions.pdf
Accessed 8.3.2020.

Hung, Aaron Chia Yuan.2017. "A critique and defense of gamification." *Journal of Interactive Online Learning*, 15(1).

Hvannberg, Ebba Thora, Effie Lai-Chong Law and Marta Kristín Lérusdóttir. 2007." Heuristic evaluation: Comparing ways of finding and reporting usability problems". *Interacting with Computers*, 19: 225–240.

Inostroza, Rodolfo, Cristian Rusu, Silvana Roncagliolo, Cristhy Jiménez and Virginica Rusu.2012."Usability Heuristics for Touchscreen-based Mobile Devices." *Ninth International Conference on Information Technology - New Generations*. Las Vegas, NV, 2012: 662-667.

ISO-9241-11. International standards for HCI and usability. Available at: http://www.usabilitynet.org/tools/r_international.htm#9241-11 Accessed 20.4.2020 Retrieved 12.1.2020

Kamper, Robert J. 2002. ”Extending the Usability of Heuristics for Design and Evaluation: Lead, Follow, and Get Out of the Way.” *International journal of Human–Computer Interaction*, 14(3&4): 447–462.

Jeno, Lucas M., Vigdis Vandvik, Sigrunn Eliassen, and John-Arvid Grytnes.2019.” Testing the novelty effect of an m-learning tool on internalization and achievement: A Self-Determination Theory approach”.*Computers & Education*, Volume 128, 2019,398-413.

Kleiman, G. M.2000. “Myths and realities about technology in K-12 schools.“ *The Online Journal of the Leadership and the New Technologies Community*, 14: 1-8.

Keller, John M and Katsuaki Suzuki.2004. ”Learner motivation and E-learning design: A multinationally validated process.” Journal of Educational Media, 29 (3): 229-239.

Kuh George D., Ty M. Cruce, Rick Shoup, Jillian Kinzie, Gonyea R. M.2008. “Unmasking the effects of student engagement on first-year college grades and persistence”. *The Journal of Higher Education*, Volume 79, Number 5, September/October 2008: 540–563. 10.1353/jhe.0.0019

Korhonen, Hannu M. I. and Elina Koivisto. 2006. “Playability heuristics for mobile games.” *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services*, Mobile HCI 2006, Helsinki, Finland, September 12-15,

Li, Yibing, and Richard Lerner.2012. “Interrelations of Behavioral, Emotional, and Cognitive School Engagement in High School Students”. *Journal of youth and adolescence 42(1) DOI: 10.1007/s10964-012-9857-5*

Lonka, Kirsi, and Elina Ketonen.2012. “How to Make a Lecture Course an Engaging Learning Experience?” *Studies for the Learning Society*, 2(2-3), 63-74.

Lonka, Kirsi. 2012. “Engaging Learning Environments for the Future: The 2012 Elizabeth W. Stone Lecture.” *The Road to Information Literacy: Librarians as Facilitators of Learning*, IFLA Publications Series 157. Edited by Roisin Gwyer, Ruth Stubbings and Graham Walton, Berlin/Munich: De Gruyter Saur: 5-30.

Molich, Rolf, and Jacob Nielsen. (1990). “Improving a human-computer dialogue”, *Communications of the ACM* 33, 3, 338-348.

Mora, Alberto, Daniel Riera, Carina S. González, and Joan Arnedo-Moreno.2017. “Gamification: a systematic review of design frameworks”. *Journal of Computing in Higher Education*, 29(3), 516-548.

Mäyrä, Frans, Jan Looy, and Thorsten Quandt. (2013). “Disciplinary identity of game scholars: an outline”. *DiGRA '13 - Proceedings of the 2013 DiGRA International Conference: DeFragging Game Studies*, August, 2014 Volume: 7

Nelson, Theodore M. 1965. “Complex information processing: a file structure for the complex, the changing and the indeterminate.” In *Proceedings of the 1965 20th national conference (ACM ’65)*. Association for Computing Machinery, New York, NY, USA, 84–100. DOI:https://doi.org/10.1145/800197.806036

Nielsen, Jacob 1989. “Usability engineering at a discount.” In *Proceedings of the third international conference on human-computer interaction on Designing and using human-computer interfaces and knowledge based systems (2nd ed.)*. Elsevier Science Inc., USA, 394–401.

Nielsen, Jacob, and Rolf Molich.1990.“Heuristic evaluation of user interfaces.” In *Proceedings of ACM CHI'90 Conference*, Seattle, WA, 1-5 April, 249-256.

Nielsen, Jacob.1994a.“Enhancing the explanatory power of usability heuristics.” *Proceedings of ACM CHI'94Conference*, Boston, MA, April 24-28: 152-158.

Nielsen, Jacob.1994b.”Severity Ratings for Usability Problems” Available at: https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/ Accessed 28.4.2020.

Nielsen, Jacob.1994c.”How to Conduct a Heuristic Evaluation” Available at: https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/ Accessed 28.4.2020.

Nielsen-Norman Group, 2020 https://www.nngroup.com/articles/usability-101-introduction-to-usability/, retrieved 27.4.2020

Murtza, Rabia, Stephen Monroe, and Robert J. Youmans. “Heuristic Evaluation for Virtual Reality Systems.” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, no. 1: 2067–71. doi:10.1177/1541931213602000

Norman, Don. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, 2013. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/tampere/detail.action?docID=1167019. Created from tampere on 2020-03-31 13:13:13.

Oh, Jeeyun, Saraswathi Bellur, and S. Shyam Sundar. “Clicking, Assessing, Immersing, and Sharing: An Empirical Model of User Engagement with Interactive Media.” *Communication Research* 45, no. 5 (July 2018): 737–63. doi:10.1177/0093650215600493.

Paavilainen, Janne. 2010. “Critical review on video game evaluation heuristics: social games perspective.” In *Proceedings of the International Academic Conference on the Future of Game Design and Technology (Futureplay ’10)*. Association for Computing Machinery, New York, NY, USA: 56–65. DOI:https://doi.org/10.1145/1920778.1920787

Paavilainen, Janne.2017. “Kokoa-evaluointimenetelmän arviointi”. Evaluation report commissioned by Kokoa Agency Oy, 28.6.2017. (Not available publicly)

Pechenkina, Ekaterina, Daniel Laurence, Grainne Oates, Daniel Eldridge and Dan Hunter.2017. “Using a gamified mobile app to increase student engagement, retention and academic achievement”. *International Journal of Educational Technology in Higher Education*, 14(1), 31.

Pearce, Jon, Mary Ainley, Mary, and Steve Howard.2005. “The ebb and flow of online learning.” *Computers in Human Behavior*. 21: 745-771.

Piccoli, Gabriele, Rami Ahmad, and Blake Ives. "Web-Based Virtual Learning Environments: A Research Framework and a Preliminary Assessment of Effectiveness in Basic IT Skills Training." *MIS Quarterly* 25, no. 4 (2001): 401-26. Accessed April 27, 2020. doi:10.2307/3250989.

Pinelle, David, Nelson Wong, Tadeusz Stach, and Carl Gutwin. 2009. “Usability heuristics for networked multiplayer games.“ In *Proceedings of the ACM 2009 international conference on Supporting group work (GROUP ’09)*. Association for Computing Machinery, New York, NY, USA:169–178. DOI:https://doi.org/10.1145/1531674.1531700

Proulx, Jean-Nicolas, Margarida Romero, Sylvester Arnab.2017. ”Learning mechanics and game mechanics under the perspective of self-determination theory to foster motivation in digital game based learning.” *Simulation & Gaming*, 48(1): 81-97.

Ryan, Richard M., and Deci, Edward L.(2000a). “Intrinsic and extrinsic motivations: Classic definitions and new directions. “ *Contemporary educational psychology*, 25(1): 54-67.

Ryan, Richard M., and Deci, Edward L (2000b). “Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.” *American Psychologist*, 55(1): 68-78. doi:10.1037/0003-066X.55.1.68

Rusu, Cristian, Silvana Roncagliolo, Virginica Rusu, and C. Collazos. 2011. “A methodology to establish usability heuristics”, In *Proceedings of 4th International Conferences on Advances in Computer-Human Interactions (ACHI 2011)*, IARIA: 59-62.

Salen, Katie, and Eric Zimmerman.2004. *Rules of Play, Game Design Fundamentals*. Cambridge, MA: The MIT Press.

Sheldon, Kennon M., Andrew J. Elliot, Youngmee Kim, and Tim Kasser.2001. "What is satisfying about satisfying events? Testing 10 candidate psychological needs. "J*ournal of Personality and Social Psychology*, 80(2): 325-339. doi:10.1037// O022-3514.80.2.325

Shernoff, David. J. 2013." Engagement as an Individual Trait and Its Relationship to Achievement". In *Optimal Learning Environments to Promote Student Engagement*. New York, NY: Springer: 97-126.

Sim, Gavin, Janet C. Read, and Phil Holfeild.2008. "Heuristics for Evaluating the Usability of CAA Applications", In *Computer Assisted Assessment Conference Proceedings,* Loughborough, England: 283-294.

Squires, D. and J. Preece.1999."Predicting quality in educational software." in *Interacting with Computers,* vol. 11, no. 5: 467-483.

Svanum, Soren., and Silvia Bigatti.2009. "Academic course engagement during one semester forecasts college success: engaged students are more likely to earn a degree, do it faster, and do it better". Journal of College Student Development 50(1):120-132.

Tondello, Gustavo F, Dennis L. Kappen, Elisa D. Mekler, Marim Ganaba, and Lennart E. Nacke. 2016. "Heuristic Evaluation for Gameful Design". In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts (CHI PLAY Companion '16).* Association for Computing Machinery, New York, NY, USA:315–323. DOI:https://doi-org.libproxy.tuni.fi/10.1145/2968120.2987729

Tondello, Gustavo, Dennis Kappen, Marim Ganaba, and Lennart Nacke.2019. "Gameful Design Heuristics: A Gamification Inspection Tool". 10.1007/978-3-030-22646-6_16.

Wan, Zeying; Yulin Fang, and Derrick Neufeld 2007. “The Role of Information Technology in Technology-Mediated Learning: A Review of the Past for the Future”. J*ournal of Information Systems Education*; Summer 2007; 18, 2; Research Library: 183

Weiser, Paul, Dominik Bucher, Francesca Cellina, and Vanessa De Luca.2015.” A Taxonomy of Motivational Affordances for Meaningful Gamified and Persuasive Technologies.” 10.2991/ict4s-env-15.2015.31.

Wharton, Cathleen, John Rieman, Clayton Lewis and Peter Polson.1994. “The cognitive walkthrough method: A practitioner’s guide.” In *Usability inspection methods*. Edited by Nielsen, J., and Mack, R. New York, NY: John Wiley & Sons

Whitton, Nicola. 2010. “Game engagement theory and adult learning”. *Simulation & Gaming*, 42, 596-609. doi:10.1177/1046878110378587

Wilson, Douglas and Miguel Sicart. 2010. “Now It’s Personal: On Abusive Game Design”. Future Play 2010. Available at: https://www.miguelsicart.net/publications/Abusive_Game_Design.pdf Accessed 20.4.2020.

Zhang, Ping.2008. “Motivational Affordances: Fundamental Reasons for ICT Design and Use.“ *Communications of the ACM*. 51: 145-147.