Research Article

Valentin Lohmüller*, Daniel Schmaderer, and Christian Wolff

A Heuristic Checklist for Second Screen Applications

https://doi.org/10.1515/icom-2019-0003

Abstract: This paper presents domain-specific heuristics for second screen applications and the development of a heuristics checklist to enable a more intuitive and structured application of the created heuristics. The heuristics presented were developed on the basis of Nielsen [12] *Ten Usability Heuristics* in a research-based approach using specific literature and a focus group. In order to evaluate the quality of the derived checklist, a heuristic evaluation of a second screen application with five users was carried out and its results compared to a user study with 20 participants. This resulted in an average validity of 0.5 and a high completeness of 0.74. The harmonic mean of these values results in an F-measure of 0.6 with an equal weighting. This value speaks for a sufficient validity of the created heuristic checklist in the first iteration.

Keywords: Second screen, heuristic evaluation, heuristic checklist, companion screen, Human Centered Design, Usability

1 Introduction

The aim of heuristic evaluations is to record the current state of software using rules, so-called *heuristics*, with the goal of improving the usability of the object under investigation. This process must be seen iteratively, so that the usability of an application increases continuously from an early stage of development ([5], p. 46). Heuristic evaluations are regarded as particularly efficient and cost-effective methods for determining usability problems and are often based on Nielsen's *Ten Usability Heuristics* [12]. In order to generate as complete a list of usability problems as possible in a particular system, it makes sense to use an adapted set of heuristics for the respective domain ([10], p. 183). Examples for already adapted heuristics are augmented reality applications [5], information appliances [1] or game design [15]. Heuristics[1] and guidelines[2] already exist in the area of second screen and smart TV, but no statement is made about their quality in the respective works. This quality is conventionally measured in *validity* and *thoroughness* of the heuristics, which refers to the correct prediction of serious usability problems of the object of investigation and the amount of problems identified by the heuristics, which are aspects addressed in the work presented here.

One of the most criticized aspects of heuristic evaluation is the loose and unstructured evaluation process ([10], p. 182) and the different interpretation of general formulated heuristics by users ([1], p. 277). In order to counteract these problems and to keep the scope of more precise heuristics manageable, a checklist was developed for the heuristics for second screen applications, which contains concrete and concise instructions for the user in order to enable efficient and comprehensive identification of usability problems ([13], 249f.).

This paper is structured as followed: the next section describes the subject area of the new heuristics *second screen* and gives on overview on *heuristic evaluation* in this context. Afterwards, the development process of the research-based first level heuristics and the derivation of the checklist points are described in more detail, and the completed heuristics are presented in section 4. Section 5 describes the evaluation of the developed heuristics, before finally, a conclusion is presented in section 6.

***Corresponding author: Valentin Lohmüller,** Media Informatics Group, University of Regensburg, Regensburg, Germany, e-mail: Valentin.Lohmueller@sprachlit.uni-regensburg.de
Daniel Schmaderer, Christian Wolff, Media Informatics Group, University of Regensburg, Regensburg, Germany, e-mails: daniel.schmaderer@stud.uni-regensburg.de, christian.wolff@ur.de

[1] Mosqueira-Rey, Alonso-Ríos, Prado-Gesto, and Moret-Bonillo [11]; Solano et al. [18].

[2] Weber, Mayer, Voit, Ventura Fierro, and Henze [19]; Pagno, Costa, Guedes, Freitas, and Nedel [14].

2 Heuristic Evaluation in the Context of Second Screen Applications

This chapter gives an overview of second screen as heuristic domain and the origins and application context of *heuristic evaluation* and how the heuristic checklist developed here represents an extension of the original concept.

2.1 Overview on Heuristic Evaluation

The evaluation of design concepts from early stages on is an essential activity in a human-centered design process and aims after its definition for the following goals (ISO [8], Evaluating the design):

- Collect new information about user needs.
- Provide feedback on strengths and weaknesses of the design solution from the user's perspective, in order to improve the design.
- Assess whether user requirements have been achieved.
- Establish baselines or make comparisons between designs.

This process is to been seen iterative and should be performed from the earliest stages on, in order to create software that meets the user's needs. This process and the positioning of the created heuristics is shown by Figure 1.

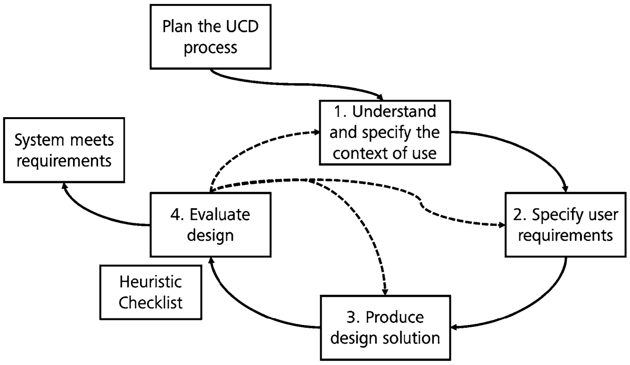
Plan the UCD process
1. Understand and specify the context of use
System meets requirements
4. Evaluate design
2. Specify user requirements
Heuristic Checklist
3. Produce design solution

Figure 1: Positioning of the created heuristic checklist in the user-centered design process.

Evaluation approaches can be either *formative* or *summative*. Formative evaluation is done during development to improve a design, and summative evaluation is conducted after the development to assess a design. Technically, usability evaluation methods can be used for both, but convention is to limit the term to formative approaches, due to the main goal to determine and resolve usability problems iteratively, before actual users are confronted with them ([6], p. 149).

The de facto standard in usability evaluation methods is a laboratory-based usability test with actual or potential users ([6], p. 151). These usability tests examine the completion of tasks within the design solution and the problems that occur while solving them, but not the users' opinion, which is obtained in user surveys. However, *evaluation by users* is not always practical or cost-effective at every stage in a design process. In this circumstances, design solutions can also be evaluated in others ways, such as in an *inspection-based* approach (ISO [8], Evaluating the design). Inspection-based evaluation describes methods where evaluators, mostly experts, examine usability-related aspects, for example cognitive or pluralistic walkthroughs, or the most distributed method in this approach heuristic evaluation, which is further elaborated in the following ([10], p. 180; [17], p. 219). Figure 2 shows the two different approaches for usability evaluation, with its most distributed methods.

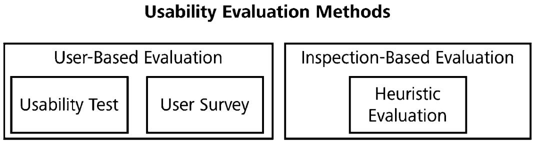

Figure 2: The two different approaches for usability evaluation, user- and inspection-based evaluation, with most distributed methods.

Heuristic evaluation is an informal usability evaluation approach that was introduced by Nielsen and Molich in 1990. In this approach evaluators produce lists of usability problem by inspecting a user interface freely and noting deviations from accepted usability principles, so called heuristics ([10], p. 180). Each problem is documented, including the violated heuristic and enough context to help understand the problem, and assigned a severity rating ([17], 216f.) The evaluation is ideally performed by usability experts, who base their judgment on previous experience and existing standards, and is repeated by multiple experts to reduce individual bias (ISO [8], Inspection-based evaluation). Heuristic evaluation is considered as cheap, fast, and easy to use, while achieving a satisfactory result and is therefore also references as discount usability method ([12], p. 25).

The concept of a free-form evaluation with a list of usability heuristics by Nielsen and Molich, was later adapted by Sears [17] to a more structured technique, the *heuristic walkthrough*. This derivation combines aspects from heuristic evaluations and cognitive walkthroughs, and consist of guided phase with prioritized list of users tasks, a list of usability heuristics, and a free exploration phase of the system ([17], p. 219). The here introduced heuristic checklist for second screen applications (cf. 4.2) combines aspects from these two approaches. It provides more structure than a heuristic evaluation and involves less effort than a heuristic walkthrough, because it only consists of one phase and does not need generated tasks.

The first set of heuristics by Nielsen and Molich [13] originated from the need to cut down the complexity of evaluating user interfaces, caused by the high number available of guidelines, which were time-consuming and difficult to use ([10], p. 180), similar to the motivation in this work for the extension to a heuristic checklist. Nielsen and Molich [13] derived the original list of usability heuristics by their understanding of typical problem areas and an informal consideration of existing guidelines. The first set compassed nine usability heuristics, the last of the following, help and documentation, was added later as tenth in 1991 ([12], p. 29):

1. Simple and natural dialogue
2. Speak the user's language
3. Minimize user memory load
4. Be consistent
5. Provide feedback
6. Provide clearly marked exits
7. Provide shortcuts
8. Good error messages
9. Prevent errors
10. Help and documentation

Nielsen later performed a more formal study, which included 101 usability principles, including the set listed above, to evaluate eleven interactive systems. The seven factors with the most explanatory power for the most usability problem formed the basis for the revised set of heuristics, to which Nielsen added three heuristics based on his own experience ([10], p. 180). The result is a revised set of usability heuristics, which is widely used among literature and also states the basis for the heuristics for second screen application introduced in this work ([12], p. 30):

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

The original heuristic evaluation was developed and applied mainly for single user, productivity-oriented desktop program, which were the majority of computer applications in the 1990s. Computer technologies have become more integrated into everyday life and versatile since then, to that degree that Nielsen´s ten heuristic may not be able to cover all usability issues in modern systems ([10], p. 183). To achieve the best possible result in heuristic evaluation in one of the diverse domains of computer systems available, it is therefore recommended to use an adapted set of heuristics. Examples of this are heuristics for augmented reality applications [5], information appliances [1], or game design [15]. Heuristics[3] and guidelines[4] exist in the areas of second screen and smart TV as well, but no statement is made about their validity in the respective studies. These works are discussed further in section 2.2. The development of heuristics for second screen applications also follows the recommendations of Ling and Salvendy [10], who encourage the development and refinement of more domain-specific heuristics to create more precise and relevant evaluation results, and by that improve the usability of that domain.

A general disadvantage of heuristics and similar principles is their high degree of abstraction, which results from the universal and vague formulation and allows a number of different interpretations by the evaluators ([1], p. 277). To compensate for these different interpretations, the heuristics can be further concretized or extended to the second-level heuristics by a more thorough formulation or by adding instructions from design guidelines ([10], p. 186). These more detailed descriptions, on the other hand, make the application of the heuristic evaluation less manageable and increases the cognitive burden on the evaluators. In order to give precise instructions, to enable a low-effort evaluation, and to keep the results consistent between different evaluators, the heuristics can be formulated in form of checklist ([5], p. 51), which was done in this work and is elaborated in the following.

[3] Mosqueira-Rey et al. [11]; Solano et al. [18].

[4] Weber et al. [19]; Pagno et al. [14].

2.2 Second Screen as Heuristic Domain

Second screening refers to the use of a second screen, such as a smartphone or tablet, while using a primary screen, such as a television ([4], p. 228). A precise assessment of how often and for how long a second screen is used is difficult, but studies assume that a large proportion – between 57 % ([3], p. 410) and 83 % ([9], p. 381) – of users at least rarely use a smartphone, tablet or laptop in parallel with a primary screen. With the advent of smart TVs, streaming sticks, set-top boxes and similar devices in recent years, second screening has gained a further perspective in the form of directly connected and bidirectional communicating second screen applications. Prominent representatives are applications such as *Netflix*, *Amazon Prime Video* or *YouTub*e, which enable users to display additional information via their mobile device and control the content on their television sets. This allows the advantages of both screens to be combined in a single second screen application.

In the development of second screen applications, numerous special features compared to conventional applications must be taken into account in order to create a satisfying user experience, such as the sensible distribution of information and responsibilities of both components or the directing of attention through notifications. Works like Mosqueira-Rey et al. [11] or the related subject area at Solano et al. [18] have already designed domain-specific heuristics according to an evaluation-based approach. In addition, there are recommendations for the design of second screen applications such as Pagno et al. [14] or Weber et al. [19], which, like Nielsen's *Ten Usability Heuristics*, can provide a basis for the derivation of new heuristics ([7], p. 226). The quality of the guidelines has not been checked in any of the named studies, which is why they are synthesized here into a heuristic checklist for which validity, thoroughness and effectiveness were determined.

3 Development of the Heuristics

According to Ling and Salvendy ([10], p. 186), domain-specific heuristics can be developed following to two different approaches: the *evaluation-based* and the *research-based* approach. In an evaluation-based approach, general usability problems with certain systems are categorized in heuristics. This type of derivation is based on empirical observations, often dependent on the examined object and therefore less suitable for the derivation of generic heuristics ([1], p. 278). The research-based approach identifies requirements and key factors of a specific domain based on appropriate literature. This method is similar to Nielsen's approach to the original heuristics, which were also synthesized from a number of existing guidelines, and thus represents a research-based approach ([7], p. 226). This approach was chosen in order to develop a domain-specific heuristic based on existing literature in the field of heuristics and guidelines in the field of second screen.

The literature used to extend Nielsen's set of usability heuristics to second screen applications is divided in existing heuristics and guidelines. Mosqueira-Rey et al. [11] formulate heuristics based on an evaluation of a single second screen application. While the resulting six heuristics contain valuable aspects, the transferability on other applications and general validity are not investigated at all. The usability heuristics by Solano et al. [18] are intended for the evaluation of interactive digital television, which is considered as closely related to the subject of second screening. The 14 resulting heuristics were created by the authors understanding of characteristic of the targeted domain and a categorization process, based on Nielsen's set. Again, no validation of the created heuristics was carried out in this work, although it is considered mandatory in the literature [6, 16, 17].

Guidelines typically contain more concrete instructions than heuristics, which is why they were used in the second step of the development process to create the checklist. The guidelines by Pagno et al. [14] were created as a summary of a series of experiments on dynamic second screen applications, but were not evaluated in the paper. Weber et al. [19] derived design guidelines for notifications on smart TVs based on their findings from a series of focus groups, an online survey, and a controlled lab study. Smart TV notifications are a central aspect in the first screen design and in many aspects transferable to the second screen, and therefore relevant. The guidelines are the result of various studies, but no statement is made about their quality.

All of the mentioned literature is assigned to the evaluation-based approach, which categorized empirical problems. The described findings hold valuable insights, but lack generalization and validation, which is why they were chosen in this paper as a supplement to the resulting heuristics.

The development of heuristics for second screen applications included the following steps: A set of existing heuristics was chosen as a basis, which was adapted and supplemented with the help of appropriate literature ([10], 183ff.). In order to keep subjective influences as low as possible, a focus group of experts from the target area was

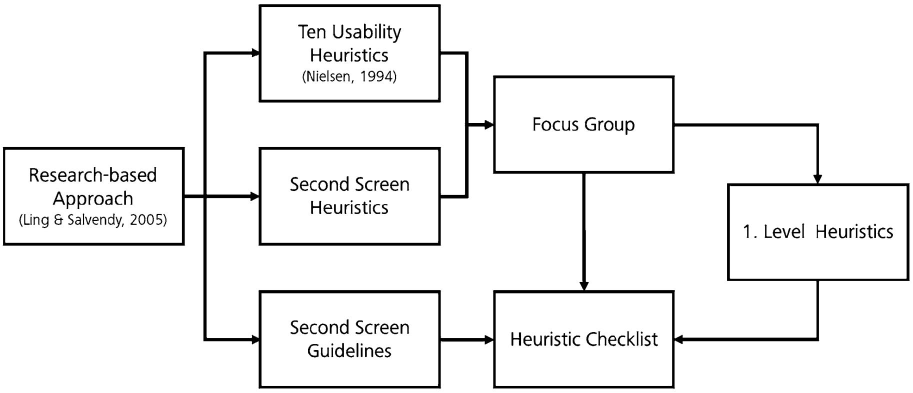

Figure 3: Procedure for the development of the heuristic checklist for second screen applications.

formed. The result represents the first level of heuristics, which was further concretized with additional literature to the heuristic checklist. Figure 3 shows the sequence of steps in the development of heuristics.

3.1 Research-Based Derivation of the First-Level Second Screen Heuristics

The *Ten Usability Heuristics* by Nielsen [12] form the basis of the adapted heuristics due to their high distribution and extensively use [16]. These were supplemented by the already existing domain-specific heuristics of Mosqueira-Rey et al. [11] and Solano et al. [18]. This process was supported by a focus group of two experts, with two and a half years' experience in the field of second screen development, to identify named and semantic duplicates, group topics, find less relevant heuristics in the field of second screening, and to extend the new heuristics by domain-specific points. Eight of Nielsen's original heuristics were found to be transferable and adapted to the new domain, and two new heuristics specifically for second screen applications were added. The complete list of heuristics for second screen applications is presented in section 4.1. It was taken care, not to exceed the total number of ten heuristics, following the example of Nielsen [12]. Although a higher number of heuristics potentially identifies more usability problems, they represent a higher cognitive burden for the user, which is why the total number of heuristics should not become too large ([10], p. 192; [13]). This set forms the first-level of heuristics for second screen applications and are formulated more comprehensively than the original set due to their specialization. Therefore, and to counteract the weakness of lack of structure in conventional heuristic evaluation ([10], p. 182), a heuristic checklist with concrete and precise instructions was developed in the second stage to facilitate the use of heuristics.

3.2 Derivation of the Heuristic Checklist for Second Screen Applications

For the creation of the heuristic checklist, individual checklist items were generated from the existing heuristics[5] and guidelines[6] in the target domain. These items were formulated as precisely as possible and referred to single aspects identified in literature. A total of 66 points were created, of which 51 were then incorporated into the previously created first-level heuristics for better overview. Due to their concrete and practical nature, guidelines are well suited for extending heuristics to be more precise ([1], p. 277), which is why they were combined with the previously gained findings from the focus group. The result of the heuristic checklist for second screen applications is described in section 4.2.

4 Heuristics for Second Screen Applications

Based on Nielsen's [12] Ten Usability Heuristics, a set of domain-specific heuristics for second screen applications was created in a research-based approach with the help of a focus group and specific literature. This set is more comprehensive than the original heuristics due to its specialization, which is why a heuristic checklist was derived from it to increase the manageability and given structure

[5] Mosqueira-Rey et al. [11]; Solano et al. [18].

[6] Weber et al. [19]; Pagno et al. [14].

in the evaluation process. For this purpose concrete checklist items were generated from literature and classified into the previously created first-level heuristics. In the following the set of first-level heuristics for second screen applications (4.1) and the heuristic checklist (4.2) are presented:

4.1 First-Level Heuristics for Second Screen Applications

1. Visibility of system status: The system should always keep the user up to date by providing appropriate feedback in a reasonable time. The user should have an overview of the current connection status between first and second screen at all times. The current content on the first screen should always be visible on the second screen to give the user a good overview of both parts of the application. Both parts of the system display the same status.
2. Match between system and the real world: A second screen application should speak the language of the user. Words, phrases and the concept of the second screen application should be presented in a natural order. If this is not the case, difficulties in using the application will increase. Especially when connecting the second screen application with the TV, complicated technical terms can make the operation more difficult or even impossible.
3. User control and freedom: The user should always have control over the content of the first screen when connected.
4. Consistency and standards: The design and layout of the interface as well as the user interaction should be consistent on both screens. In addition, standardized icons, conventions, and terminology should be used.
5. Error prevention: The design and explanation of a second screen application should prevent the occurrence of errors as far as possible. If errors do occur, it is important to describe them as clearly and concisely as possible in order to make it easier for the user to handle the error messages.
6. Recognition rather than recall: Objects, options and actions should be visible and recognizable in second screen applications. The user should be aware of his possibilities in all areas and not have to remember them.
7. Aesthetic and minimalist design: A second screen application should not occupy the user with irrelevant information as it distracts him from the relevant information. This is especially important when the system is communicating with the user. Notifications should be subtle and not too frequent. Effects and animations should be used with care to avoid distracting the user.
8. Help users recognize, diagnose, and recover from errors: If errors occur in a second screen application, the help should be formulated in the user's language. In the case of errors that the user can correct by himself, the error message should be accompanied by instructions for correcting the error. In the case of errors that the user cannot fix, this should be clearly stated. The cause of the error should always be clear, especially in relation to the connection process.
9. Connection process: The connection process should be as simple as possible and available from anywhere. The first and second screens should be assigned the correct roles, with control on the second screen and media presentation on the first screen.
10. Use a second-screen when it adds value: A second screen application should only be used if it provides added value for users.

The adapted heuristics are intended for the same use as the original set of heuristics. It can serve as a basis for heuristic evaluations and walkthroughs and draw the evaluators' attention to important aspects regarding second screen applications. A known weakness of conventional heuristic evaluations is the unstructured process, supported only by the sometimes vague formulations of the heuristic ([10], p. 182). In order to counteract this problem and to facilitate the evaluation process, the first-level heuristics were extended to a checklist, which is presented below.

4.2 Heuristic Checklist for Second Screen Applications

1. Visibility of the System Status
   a. Does the application give the user feedback?
      i. At performing key actions
      ii. At reasonable time
   b. Is the status of the connection kept updated?
   c. Are the screens keep synchronized instantaneously?
2. Match between the system and the real world
   a. Does the application speak the user's language?
      i. Understandable terms/descriptions
   b. Does the application show the information in a natural order?
   c. Does the sequence of activities follow the user's mental processes?
3. User control and freedom

   a. Is the navigation simple and intuitive for the operating system?
      i. Menu
      ii. Search bar
   b. Does the application provide different options?
      i. Return to top level
   c. Is the user able to explore the application freely?
   d. Is the user able to control the content of the TV at any time?
4. Consistency and standards
   a. Does the application follow the design guidelines of the using platform?
   b. Is the consistency between the two applications given?
      i. Terminology
      ii. Controls
      iii. Graphics/Icons
      iv. Focus on one guideline (if multiple apply)
5. Error prevention
   a. Is there a help for novice users?
   b. Does the application provide appropriate error messages?
6. Recognition rather than recall
   a. Is the relationship between the controls and their actions obvious?
   b. Does the user know what options he has and where to go?
      i. Main elements of application always available
      ii. Help available if needed
7. Aesthetic and minimalist design
   a. Does the application only show relevant and necessary information to the user?
      i. Titles and headlines short but descriptive
   b. Is the application design appropriate?
      i. Distance between elements
      ii. Size
      iii. Placement
   c. Are the elements of the application visible?
      i. At the visual range of watching TV
      ii. At various types of lighting
   d. Are notifications subtle and not to frequent?
      i. At least 30 seconds apart?
8. Help users recognize, diagnose and recover from errors
   a. Does the application provide clear messages with indicating errors and solutions for errors?
      i. Connection error, application crashes, etc.
   b. Are the error messages written in an accurate way?
      i. Not blaming the user
      ii. Non-derisory tone
   c. Does the application provide users a clear and simple help, in their own language?
9. Discovery and Connection
   a. Is the pairing of main and secondary display simple and intuitive?
      i. 1-3 clicks needed
      ii. Direct Response after pairing
   b. Is the separation between the two applications and devices clear?
   c. Is the main logic on the mobile device?
   d. Is the main content shown on the first screen?
10. Use a second-screen when it adds value
   a. Does the second screen add value to the first screen?
   b. Does the second screen improve the content navigation?
   c. Does the second screen give the user a better user experience?

The level of structure provided by the checklist is between a heuristic evaluation, where guidance is only provided by the set of used heuristics, and a heuristic walkthrough, which consists of a phase in which task completed and a second phase in which the application is examined freely ([17], p. 219). This approach requires more preparation and tends to take more time during execution, which contradicts the low-effort character intended by heuristic evaluation. Therefore, the heuristic checklist is seen as a compromise that combines aspects of both approaches. The checklist is intended to be worked through systematically by the evaluators. The individual checklist elements are partly aimed at the general workflow, which encourages the evaluators to explore the application freely, and in part at the closer inspection at important aspects, such as the connection process or error handling. This allows a mixture of free-form evaluation and guidance during evaluation, which led to much positive feedback by the experts in the evaluation. This evaluation is part of the validation process of the here introduced heuristic checklist and is presented in the next section.

5 Validation of the Heuristic Checklist

The derivation of new heuristics generally consist of two steps: *heuristic development* and *heuristic validation*. All the literature from the second screen area that was used for the development of the heuristics have omitted this second step, although it is considered very important, which is why a first validation of the heuristic checklist was carried out this work. In the validation phase, the newly developed heuristics are typically compared with Nielsen's original set by conducting empirical studies or benchmarked with user testing results. The adapted set of heuristics is usually more effective than the original set because it fits the evaluated domain better, which makes this approach less meaningful ([10], 183ff.). Therefore, a validation consisting of a heuristic evaluation, with the developed checklist, with the comparison of user tests was chosen.

5.1 Validation Methodology

Conventionally three measures are used for the evaluation of heuristics: *validity*, *thoroughness*, and *reliability* ([1], p. 281; [6], p. 160; [10], p. 187; [17], p. 214). The formulas used for the calculation of these measures will be presented in the next section during their application; in the following general concepts are briefly discussed.

Validity describes the ratio of the real problems found to all identified problems, thereby describing the correctness of the identified problems by the evaluators. This measure is based on the concept of *precision*, used to describe information retrieval performances. It is also based on the belief that evaluators are able to identify issues as usability problems that are not actual problems. It can be argued, that any problem identified by users or experts is a problem worth further investigation and thus cannot be *false*, which contradicts this understanding of validity. This discussion remains controversial, but goes beyond the scope of this work. Nevertheless, validity is a standard measure for the comparison of interface evaluation techniques and holds value with correct interpretation ([17], p. 214)

Thoroughness indicates how many of the predicted problems are actually found, is perhaps the most attractive measure and is based on the concept of *recall*. Similar to the calculation of recall the determination of the denominator is problematic, since it is difficult to know how many problems exist in total. Commonly, the sum of all identified problem or all problems encountered by users are used, because these are found to be real, although this may not a perfect estimate ([6], p. 161; [17], p. 215).

Reliability is a measure of consistency of testing results across different evaluator. There are various approaches to calculating reliability, such as *Pearson's r* [12], *Cohen's kappa*, the *ratio* of standard deviation of the number of problems to the average number of problems found [17], or *Kendall's coefficient* [12]. Although it is usually desirable to achieve constant results among different evaluators, the total result of the group is relevant, not the ones by single experts. If individual evaluators find completely different results and they are all relevant, this variety should be encouraged ([6], 167f.).

As mentioned before, validity and thoroughness are measures based on precision and recall, and are not sufficient on their own to make a statement about the overall effectiveness of the applied heuristic. This *effectiveness* is usually defined as the product of thoroughness and validity. Due to the quadratic relationship, the calculated result is strongly influenced by a low value, and represents the relation rather unsatisfactorily. Hartson et al. ([6], p. 166) therefore describe analogous the precision and recall the calculation of a weighted harmonic mean, the *F-measure*.

In order to calculate these values, a second screen application was evaluated by five experts with the help of the heuristic checklist and the collected results were compared with those from a usability study of 20 potential users. The object of investigation was a second screen application that was still under development. The aim was to identify problems regarding usability at an early stage before they affect end users. For a meaningful evaluation of reliability, further heuristic evaluations of second screen applications would have to be carried out, which was not possible in the first iteration, and therefore this measure was not calculated. Nielsen and Molich ([13], p. 255) recommend between three and five users of a heuristic evaluation for the most efficient determination of usability problems. The experts examined the application with the help of the checklist, classified found problems with regard to the heuristics and assigned a severity level according to Nielsen [12] of 0, no problem, to 4, must be solved. Finally, the experts were asked about the heuristics used.

The usability problems in the application predicted in the heuristic evaluation were used to generate the tasks of the user study, in order to increase the power of user testing for exposing all predicted problems that really exist ([7], p. 227). For this purpose, semantic and content-related duplicates in the problems found were removed and grouped thematically. From the resulting groups, feedback, help, error, connection, search, menu, video, navigation and playlist, tasks were created for the users that they are not directly confronted with the corresponding problems, but all problem areas were examined. The aim was to check to what extent the predicted problems correspond to the heuristic evaluation of real user problems. Finally, the System Usability Scale (SUS) [2] of the examined application was surveyed and a partially structured survey was carried out. Figure 4 gives a schematic

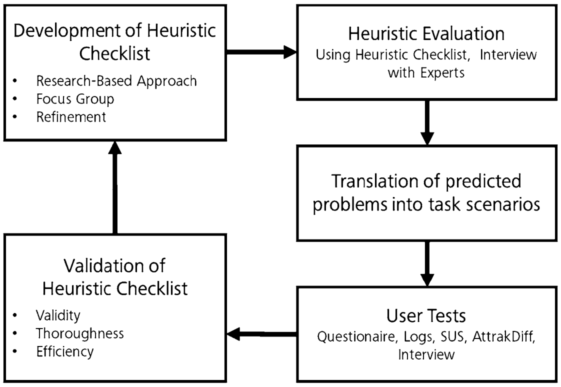
Development of Heuristic Checklist
• Research-Based Approach
• Focus Group
• Refinement
Heuristic Evaluation
Using Heuristic Checklist, Interview with Experts
Translation of predicted problems into task scenarios
User Tests
Questionaire, Logs, SUS, AttrakDiff, Interview
Validation of Heuristic Checklist
• Validity
• Thoroughness
• Efficiency

Figure 4: Schematic Overview of the Validation Process of the Created Heuristic Checklist.

overview of the applied validation process of the created heuristic checklist. In the following, the first iteration of the process is further elaborated.

5.2 Results of First Validation Iteration

The here described iteration should be seen as the first step in a thorough validation, as further second screen applications and heuristic evaluations can be carried out in order to gain even better insight into the here developed heuristics. The content of the heuristics can also be further adapted or extended in future work. In the following, the measures, *validity*, *thoroughness*, *efficiency* and *f-measure*, calculated from the first iteration in the validation process are described:

Validity

Heuristics have a high validity if as many as possible of the predicted problems from the evaluation match the real/actual problems of the users ([6], p. 163f.). The real/actual problems found are seen as the intersection of the problems found by users and experts.

$$\text{Validity} = \frac{\text{number of real problems found (experts} \cap \text{users)}}{\text{number of issues identified as problems (sum experts)}}$$

$$\text{Validity} = 14 \div 28 = 0.5$$

Of the 28 predicted problems, 14 were confirmed by users, resulting in a validity of 0.5 of the applied heuristics for the object under investigation. This value indicates a mediocre validity and suggests that rather different errors or a different amount was found between users and experts. The validity of a heuristic usually decreases with the number of evaluators, which is at the upper end of the suggested size in this study. Furthermore, the research object was still under development, which could be a reason for the rather high number of usability problems. Problems that have been predicted by experts and not confirmed by the users nevertheless add value to the development of an application because these errors can be eliminated early on. The controversy of interpretation of false error in the context of validity was outlined in the previous section. Nonetheless, the high number of errors found by the experts outside the applied measures can be seen as positive for the development process, since this is where the real meaning of a heuristic evaluation lies: the efficient and cost-effective identification of usability problems before actual users are confronted with them.

Thoroughness

Thoroughness describes the number of existing problems that could be identified by the heuristic evaluation. Again, the intersection of the problems found is seen by experts and users as real / actual problems found and the sum of all user problems as the number of real existing problems. The problems of the experts not found by users thus turned out to be false positives and are taken into account in the validity ([6], p. 163f.).

$$\text{Thoroughness} = \frac{\text{number of real problems found (experts} \cap \text{users)}}{\text{number of real problems that exist (sum users)}}$$

$$\text{Thoroughness} = 14 \div 19 = 0.74$$

With a value of 0.74, the heuristic checklist is highly complete. This is partly due to the high number of problems identified by the experts.

Efficiency and F-Measure

The effectiveness of heuristics can be calculated based on the measures thoroughness and validity. This results from a simple multiplication of the two values.

$$\text{Efficency} = \text{Validity} \times \text{Thoroughness}$$

$$\text{Efficency} = 0.74 \times 0.5 = 0.37$$

The rather low value 0.37 is due to the mediocre validity of the heuristics. Hartson et al. ([6], p. 165) notes the strong influence of a low value on this measure of effectiveness and describes a calculation of a weighted F-measure:

$$F = \frac{1}{\alpha\,(1/\text{Validity}) + (1-\alpha)(1/\text{Thoroughness})}$$

An equal weighting of both values (α = 0.5) results in an F-measure of 0.6, which describes a weighted mean between validity and thoroughness. This value lies in the middle to positive range and describes an acceptable result of the heuristics in the first iteration.

6 Conclusion

This paper introduced general heuristics and an extended heuristic checklist for the domain of second screen applications. These heuristics were derived in a research-based approach on basis of Nielsen [12] *Ten Usability Heuristics*, which were adapted and supplemented with the help of appropriate literature from the field second screening. In order to keep subjective influences as low as possible, a focus group of experts from the target area was formed, who created the first level of heuristics for second screen applications. A heuristic checklist was created with the help of further literature and the insights from the focus group, to ease the use of the heuristics and to counteract one of the most criticized aspects of heuristic evaluation, the loose and unstructured evaluation process only supported by the list of used heuristics ([10], p. 182).

To assess the quality of the developed heuristics, a heuristic evaluation with five experts and the created checklist was conducted. The predicted problems were matched with the results of a user test of the same second screen application. The results indicate a mediocre validity of 0.5 and an acceptable thoroughness of 0.74. The weighted mean of these measures results in a sufficient efficiency of 0.6 of the developed heuristic checklist for the first iteration.

This work is to be seen as the first step of a thorough validation of the heuristics produced here. This process can be extended to other second screen applications and heuristic evaluations in order to gain even better insight into the here presented heuristics. The content of the heuristics can also be further adapted and extended. The concept of heuristics in the form of concrete and concise key points was perceived by the evaluators as particularly positive, which can be investigated in further studies.

Funding: We thank the Bavarian Ministry of Economic Affairs and Media, Energy and Technology (Grant Nr. IUK475/002), which contributed funding to this research, as well as our project partner *MEKmedia* for its support in carrying it out.

References

[1] Böhm, P., Schneidermeier, T., & Wolff, C. (2014). Heuristiken für Information Appliances. In A. Butz, M. Koch, & J. Schlichter (Eds.), *Mensch & Computer 2014 – Tagungsband* (pp. 275–284). Berlin: De Gruyter Oldenbourg.

[2] Brooke, J. (1996). SUS – A quick and dirty usability scale. *Usability evaluation in industry*, 189 (194), 4–7.

[3] Busemann, K., & Tippelt, F. (2014). Second Screen: Parallelnutzung von Fernsehen und Internet. *Media Perspektiven*, *7*, 408–416.

[4] Cunningham, S., & Weinel, J. (2015). Second screen comes to the silver screen: A technology feasibility study regarding mobile technologies in the cinema. In R. Picking (Ed.), *2015 Internet technologies and applications (ITA): Proceedings of the sixth international conference* (pp. 228–232). Piscataway, NJ: IEEE. https://doi.org/10.1109/ITechA.2015.7317400.

[5] Guimaraes, M. d. P., & Martins, V. F. (2015). A Checklist to Evaluate Augmented Reality Applications. In *2015 XVII Symposium on Virtual and Augmented Reality (SVR): 25–28 May 2015, São Paulo, Brazil* (pp. 45–52). Piscataway, NJ: IEEE. https://doi.org/10.1109/SVR.2014.17.

[6] Hartson, H. R., Andre, T. S., & Williges, R. C. (2003). Criteria For Evaluating Usability Evaluation Methods. *International Journal of Human–Computer Interaction*, *15*(1), 145–181. https://doi.org/10.1207/S15327590IJHC1501_13.

[7] Hvannberg, E. T., Law, E. L.-C., & Lárusdóttir, M. K. (2007). Heuristic evaluation: Comparing ways of finding and reporting usability problems. *Interacting with Computers*, *19*(2), 225–240. https://doi.org/10.1016/j.intcom.2006.10.001.

[8] ISO (2010). *Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme*. (DIN EN ISO, 9241-210): Beuth Verlag.

[9] Johnen, M., & Stark, B. (2015). Wenn der Fernseher nicht mehr ausreicht: Eine empirische Analyse der Second Screen-Nutzung [When watching television becomes insufficient: An empirical analysis of second screen usage]. *Studies in Communication and Media*, *4*(4), 364–405.

[10] Ling, C., & Salvendy, G. (2005). Extension of heuristic evaluation method: a review and reappraisal. *Ergonomia An International Journal of Ergonomics and Human Factors*, *2005* (27 (3)), 179–197.

[11] Mosqueira-Rey, E., Alonso-Ríos, D., Prado-Gesto, D., & Moret-Bonillo, V. (2017). Usability evaluation and development of heuristics for second-screen applications. In S. Y. Shin, D. Shin, & M. Lencastre (Eds.), *Proceedings of the Symposium on Applied Computing – SAC'17* (pp. 569–571). New York, New York, USA: ACM. https://doi.org/10.1145/3019612.3019883.

[12] Nielsen, J. (Ed.). (1994). *Usability inspection methods*. New York NY u. a.: Wiley.

[13] Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In J. C. Chew, J. C. Carrasco, & J. Carrasco Chew (Eds.): *Vol. 1990. Human factors in computing systems, CHI'90 Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems (pp. 249–256). Reading, MA: Addison-Wesley. https://doi.org/10.1145/97243.97281.

[14] Pagno, B., Costa, D., Guedes, L., Freitas, C. D. S., & Nedel, L. (2015). Guidelines for Designing Dynamic Applications with Second Screen. In *2015 XVII Symposium on Virtual and Augmented Reality (SVR): 25–28 May 2015, São Paulo, Brazil* (pp. 42–51). Piscataway, NJ: IEEE. https://doi.org/10.1109/SVR.2015.14.

[15] Pinelle, D., & Wong, N. (2008). Heuristic evaluation for games. In M. Burnett (Ed.), *The 26th Annual CHI Conference on Human Factors in Computing Systems, CHI 2008: Conference proceedings; April 5–10, 2008 in Florence, Italy* (p. 1453). New York, NY: ACM. https://doi.org/10.1145/1357054.1357282.

[16] Rusu, C., Roncagliolo, S., Rusu, V., & Collazos, C. (2011). A Methodology o Establish Usability Heuristics. In *ACHI 2011: The Fourth International Conference on Advances in Computer-Human Interactions* (pp. 59–62).

[17] Sears, A. (1997). Heuristic Walkthroughs: Finding the Problems Without the Noise. *International Journal of Human–Computer Interaction*, *9*(3), 213–234. https://doi.org/10.1207/s15327590ijhc0903_2.

[18] Solano, A., Rusu, C., Collazos, C., Roncagliolo, S., Arciniegas, J. L., & Rusu, V. (2011). Usability Heuristics for Interactive Digital Television. In E. Borcoci & J. Bi (Eds.), *AFIN 2011: The Third International Conference on Advances in Future Internet: August 21–27, 2011, Nice/Saint Laurent du Var, France*. Wilmington, DE, USA: IARIA.

[19] Weber, D., Mayer, S., Voit, A., Ventura Fierro, R., & Henze, N. (2016). Design Guidelines for Notifications on Smart TVs. In P. Whitney, J. Murray, S. Basapur, N. Ali Hasan, & J. Huber (Eds.), *TVX 2016: Proceedings of the 2016 ACM International Conference on Interactive Experiences for TV and Online Video: June 22–24, 2016, Chicago, IL, USA* (pp. 13–24). New York, NY: ACM. https://doi.org/10.1145/2932206.2932212.

Bionotes

Valentin Lohmüller
Media Informatics Group, University of Regensburg, Regensburg, Germany
Valentin.Lohmueller@sprachlit.uni-regensburg.de,
valentin.lohmueller@ur.de

Valentin Lohmüller has a master degree in media informatics and is a researcher in the area of second screen interaction at the media informatics group, University of Regensburg. His research interests are related to the design, development, and evaluation of second screen applications and creating insights on how to create a satisfying usability and user experience in this area. Another focus is on the cross-platform development of second screen applications.

Daniel Schmaderer
Media Informatics Group, University of Regensburg, Regensburg, Germany
daniel.schmaderer@stud.uni-regensburg.de

Daniel Schmaderer has been enrolled in the master's program in media informatics at the University of Regensburg since 2018, where he also completed his Bachelor of Arts degree. His research focuses on second screen applications, usability engineering, and HCI.

Christian Wolff
Media Informatics Group, University of Regensburg, Regensburg, Germany
christian.wolff@ur.de

Prof. Dr. Christian Wolff has been professor for media informatics at the University of Regensburg since 2003. After his Ph. D. thesis, in which he designed an interactive retrieval frontend for factual data (1994, University of Regensburg), he worked as an assistant professor at the Computer Science Department of the University of Leipzig from 1994 to 2001. In 2002, he became professor for media informatics at the Chemnitz University of Technology. As an information and computer scientist, he has a long record of research in electronic publishing, information retrieval and HCI. Currently, there is a strong focus on new interaction technologies like eye tracking or multitouch and gestural interaction in his group.