Humanist Evaluation Methods in Locative Media Design

Anders Fagerjord

Department of Media and Communication, University of Oslo

anders.fagerjord@media.uio.no

ABSTRACT

Media design can be used for research purposes if it includes a clearly defined research question and clear evaluation to see whether an answer to the research question has been found. Using a project with locative media for classical music communication as my example, I discuss common evaluation methods from the User experience field, observing that they all tend to test “interface” and not “content.” Instead, I propose three other methods of evaluation that have a basis in humanist theories such as textual analysis and genre studies. These are: (1) Qualitative interviews with evaluators after the evaluation, asking them to describe the service in their own words, followed by a semantic analysis to get at how they have understood the service; (2) within-subject A/B tests with alternative versions that are different in key aspects, and (3) peer review by experienced design researchers who are likely to have a more fine-tuned vocabulary to express their opinions.

KEYWORDS:

Media design, genre design, evaluation, design theory-humanist theory, methodology, philosophy of science

The Journal of Media Innovations 2.1 (2015), 107-22.

http://www.journals.uio.no/index.php/TJMI

INTRODUCTION

Computer Science research has provided many of the fundamental technologies for modern media, such as hypertext, bit-mapped graphics, and multimedia (Moggridge 2007). For forty years, calls have been voiced for a parallel design activity in the humanities, using humanist theories both to inform the design and to advance humanist theories on text, image, and communication (Nelson 1974; Ulmer 1989; Nelson 1992; Liestøl 1999; Bolter 2003; Moulthrop 2005; Nyre 2014). According to Hevner, March, Park, and Ram (2004), design science builds new artefacts from a "knowledge base" of foundations and methodologies, and the resulting design adds to this knowledge base (p. 80). If we accept their view, a design method for the humanities could draw on humanist knowledge of genres, storytelling, rhetoric, visual culture, and much more to create new experiences, services, and genres, in order to advance this humanist knowledge base. Design, understood as the practice of creating detailed plans for a possible future artefact, has been described as assembling known ideas in new combinations (Krippendorff, 2006). Humanities design uses concepts from theories as inspiration (Ulmer, 1989), and aspects of earlier works as described by scholars as building blocks (Liestøl, 1999). Humanities design for research needs a clearly stated research question and rigorous evaluation of the finished product to see what answers are found the research question (Fagerjord, 2012). Several methods are well established to test the usability of physical and digital products. User experience (UX), understood as a combination of usability, utility, and hedonistic quality is a discipline with a range of methods that are agreed to be useful (Hartson & Pyla, 2012). I will argue that these methods are too coarse to give insight into what is the core interest of a humanist approach: the user's experience of the text. UX methods are meant to test the success of a user's access to system data, or "content", but not to discern between different kinds of "content", such as images, stories, or music. I will then propose three alternative methods that may be used to get more nuanced insights into how users experience a service.

I base my discussion on the design project "Musica Romana", a web site for communicating classical music via mobile devices, aimed at tourists in Rome with an interest in music and possibly also in art or architecture. A JavaScript accesses the phone's geolocation sensors and determines the device's location. The result is a web page listing eight (in the current version) of Rome's historical churches (see figure 1) and the distance to each of

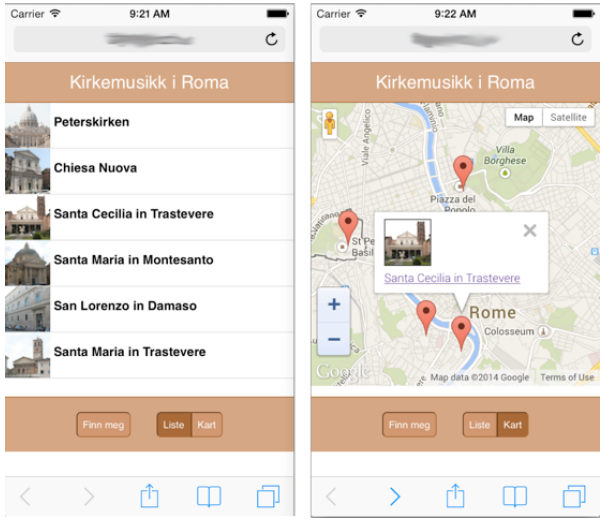

Figure *1*. Musica Romana's initial list view (left) and map view (right), listing the churches that may be visited with the service.

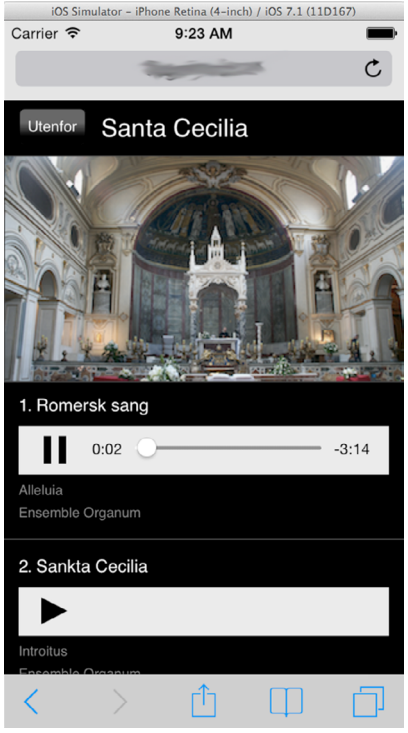

Figure 2. The Inside screen of one of the churches, where music and commentaries may be played back.

them. The churches are also drawn on a map.

For each church, there are two or three audio tracks meant to be played inside (figure 2). Music written for that church is played back and a narrator gives some details of the music's history and structure and compares it to the church's architecture and the art inside. Each commentary lasts about a minute (for more discussion of this service, see Fagerjord, 2012).

Our main focus in this project is not on technology or user interface. We use simple and well-known HTML and JavaScript functions, and although it took us several tests and iterations to arrive at the user interface, we do not claim that it is remarkable. It is the effect of the text, images and music that we want to examine: this is the what in much computer research is just called "content" - but it is the core of most media studies, film studies, art history, or comparative literature. One main topic in these fields is the concept of 'genre.' A genre is family of texts that display recurring traits (for a discussion of kinds of traits, see Altman 1984) while still having individual features, or "instances of repetition and difference", as Neale (2006) puts it. A genre invokes expectations in the audience and guides comprehension, as it signals how what rhetorical situation (Miller 1984) the text is aimed at. "Musica Romana" is what Gunnar Liestøl (2009) has called a "genre prototype," an example of a possible future genre, aimed at a certain rhetorical situation that may be recurring. This is in fact precisely what Madeleine Akrich (1992) calls a 'scenario' or a 'script': an anticipation of how users may want to use an artefact, in this case, a mobile Web service.

In order to create our script and understand the rhetorical situation, we observed tourists at the sights, noting regular behavior, and formed our scenario of how the service could be used. We also studied similar genres, such as tourist audio guides. Our prototype combines genre traits from radio (blending music and talk, explaining music via historical facts and anecdotes) and tourist guides (lists and maps of sites to visit). Our goal is to be able to argue our prototype is suitable in a recurring rhetorical situation, and that we can formulate this as a set of guidelines for this situation.

In earlier tests of our application in Rome, we used observation, "think-aloud" methods, semi-structured interviews and a simple survey, methods adapted from the Human-Computer Interaction (HCI) field. We were able to make the interface easier to use, but observational methods hardly gave any insight into how participants experienced the churches together with the music and the commentaries. To find this out, we interviewed and surveyed the users, who responded that they liked

our application. Survey scores were all positive. We were encouraged, but what could we add to the research literature, the "knowledge base"? In the explanatory audio in the application we have, for example, taken care to point out synesthetic parallels, drawing attention to structural similarities in a church's architecture and music written in the same period. Can we now conclude that synesthetic parallels are a general principle that works for situated sound? No. We can't even be certain that it worked in this case: The users may very well have liked other aspects of the application.

In the following, we will discuss the strengths and weaknesses of different evaluation methods that are used in design of experiences, media, and genres, and then describe some new approaches, based in the humanities, that we have used in the summative evaluation of our proposed new genre: semantic analysis of user interviews, within-subject A/B testing, and peer review.

TESTING GENRE DESIGN

In Klaus Krippendorff's words, "designers create and work out realistic paths from the present towards desirable futures and propose them to those who can bring a design to fruitition" (Krippendorff 2006, p. 29). While science is the study of what is, either in nature or society, design is a proposal of what can be made: "In other words, scientists are concerned with explaining an observable world, designers with creating desirable worlds, and statements about either of these worlds call for vastly different methods of validation" (p. 261).

On the other hand, disciplines such as engineering, computer science, information systems, pharmaceutics, or medicine also create artefacts belonging to desirable futures while being closely tied to science. These ties are of two kinds: First, the construction of artefacts relies on theories created by (observational) science. Second, the effectiveness of the artefact (the validity of the desirability claim) is tested using similar methods to those used to create the theories, mainly observations and statistics (March & Smith, 1995). For example, Bruno Latour and Steve Woolgar (1986), recount how advanced machinery found in a biology lab in 1976 was created within different sciences, relying on earlier results and theories in the same sciences (although in different fields).

I see two principles followed in design science that we may try as we search for humanist design methods: (1) When prototypes of a design are created, we can evaluate them with observational methods, in the same way as it is done in engineering, computer science, or medicine. (2) When designs are based in theories, we can use the same methods that were used in creating the theories to validate the designs. In genre design, these are likely to be genre theories, which are made by close textual analysis of a large number of texts. A textual analysis of the new genre could be a way of validating the design. We will discuss the two principles in turn, beginning with observational methods from the sciences, primarily from psychology.

OBSERVATION

The most basic form of evaluation of new designs is observation of use. Evaluators are asked to try out certain features of the new artefact, while members of the design team observe them. Computer applications are often tested in a usability lab, equipped with a one-way mirror additional observers can hide behind, and video cameras recording both the user's movements and what happens on the screen (Hartson & Pyla, 2012).

Locative genres such as Musica Romana can hardly be evaluated in a laboratory. It is in their very nature that they are made to be experienced in a certain place, so evaluators must be taken to the place in question. The main benefit of the proposed genre was also not the interface, but the style of presenting information. The interfaces had to be usable to be sure, and user observation, especially of critical incidents (Andersson & Nilsson, 1964; Flanagan, 1954; Hartson & Pyla, 2012) contributed much to this. But when users were able to access the information in the applications, there was very little to be learned from observing their reaction to what was presented to them. Those who tried out the Musica Romana service walked around listening, with little or no expression of whether they liked what they heard, or if they found it boring, difficult, or interesting, but too long. Other methods are needed to know what goes on in the heads of readers and listeners.

One solution to this is the so-called "think-aloud" test, or protocol analysis, where evaluators are given tasks to solve, and instructed to "think aloud" while performing the tasks, telling the observers how they think and what strategies they use to solve the questions (Lewis, 1982). It has become the most common way of testing computer interfaces, and was popularized by Jakob Nielsen and Steve Krug among others (Nielsen, 2000; Krug, 2010). According to Harson and Pyla (2012) "the think-aloud technique is also effective in assessing emotional impact because emotional impact is felt internally and the internal thoughts and feelings of the user are exactly what the think-aloud technique accesses for you" (p. 440). Krippendorff (2006) on the other hand, points out that a known limitation of this method is that many tasks are made automatically in real life, and that verbalizing them slows them down, or may even impair the respondent's ability to perform them (p. 226).

In a study that can serve as an example of this method, Nielsen and Loranger asked 69 participants a set of about 15 tasks for their usability study of a wide range of web sites (Nielsen & Loranger, 2006). Of these only six tasks can be said to concern the "content" of web sites; the information contained in text and images, the style of the prose, and so on. All of them ask for what Nielsen and Loranger call "informational value," in questions such as "list the two main causes of…" or "find out why…" The questions resemble school homework, in fact. Nielsen and Loranger do not ask respondents to evaluate aesthetic qualities or the experience of reading, yet many of the verbatim quotes from evaluators reproduced in Nielsen and Loranger's book show important insights into how readers react to texts, although they are mainly complaints about pages users do not understand or find tedious to read.

It should also be noted that Nielsen and Loranger's preferred method was comparative: They compared web pages to other web pages. Several tasks were web wide, asking evaluators to surf the net for answers. This method could not be used for the "Musica Romana" project. Like most locative web sites and other genre experiments, it is unique in its location, and we have not found simi-

lar music services to compare it with in other locations either.[1] The researcher may create alternative solutions, however, asking evaluators to think aloud while using versions that differ in important aspects, and then analyze the differences in their comments. This will be expanded below.

In the "Musica Romana" project, we did use the think-aloud method when testing the navigation system. Users were asked to use the application to locate the nearest church in the program, and to find their way there, thinking aloud when reasoning. This was a helpful technique, and we discovered several improvements to the interface from this evaluation.

Thinking aloud isn't always practical, or even possible, however. We tested our app inside churches, where continued discussion could disturb devoted church visitors. Evaluators were also listening to music and spoken commentary, and talking aloud would make it impossible to listen carefully, thus spoiling the experience they were about to evaluate.

1 Our initial research interest was in bringing music into these locations, so music was at the core of the evaluation. The most similar project we know is the "Chopin Benches" in Warszaw, Poland, but these contain mostly speech, and just short snippets of music.

ASKING THE USERS

A more indirect observation method is the survey. Distributing a survey to evaluators after they have tried out a new artefact is not an observation of their use as such, but it is a way of making their experiences observable, and, perhaps more important, quantifiable. Experiences are translated into a few categories, and frequencies in each of these categories are summed up and analysed statistically. Surveys are a way of measuring using a common yardstick, allowing for comparison between tasks.

Survey evaluation is contested within design science, however. Hartson & Pyla have contended that a "questionnaire is the primary instrument for collecting subjective data from participants in all types of evaluation" (Hartson & Pyla, 2012, p. 444). Krippendorff, however, stated bluntly that "structured interviews and questionnaires probably are the least informative methods for gaining insights ..." (Krippendorff, 2006, p. 223).

Krippendorff's critique notwithstanding, we wanted to compare our results with other studies that have used survey evaluation, so we created and distributed a simple questionnaire to our evaluators in the first round of evaluation of "Musica Romana." Evaluators were asked to rate the service by judging 11 questions on a 5-point Likert scale (see Appendix). To distribute it to only five evaluators hardly yields any statistical power to our research, but we used this as a pilot study to see if this survey was likely to give important insights.

What was most striking was that they all gave a 5 (strongly agree) to the statement "It was exciting to be present where the music was first played." We in the design team interpreted this as a strong encouragement to continue the project. There were less unison feedback to questions about the combination of music, music history, art, and architecture, but as all the users were positive towards the service, we interpreted this to mean that our matching of music and art through synaesthetic parallels worked as intended. Still, we could not get rid of a feeling that we might just be looking for indications that the users liked what we hoped they would like, and/ or that they have similar tastes as we do, and that the system we are proud of can be considered a success. But popularity is not success in research; knowledge is. An average score does little to advance our knowledge of new genres in location-based media. This experience supported Krippendorff's view on surveys; they give little insight into the actual experience of a new design.

More sophisticated surveys than ours exist. Psychologists have in recent years investigated what they call emotional impact or hedonic quality, such

as how appealing the user finds a product's look and feel. AttrakDiff is one questionnaire created to measure hedonic quality (Hassenzahl, 2000; 2001). Its creators have tested it statistically and found it valid, but remind us that while the questionnaire measures how pleasurable a product is, it cannot say what about the product that creates pleasure or indifference. It is based on a model of user experience where "appeal" is seen as the combination of "ergonomic quality" and two forms of emotional impact, called "stimulation" and "identity". The three kinds of quality, as well as the combined "appeal" are measured using semantic differentials (Osgood, Suci, & Tannenbaum, 1957). Respondents are asked to place their opinion of the product on a seven-point Likert scale between two adjectives, for example:

Pleasant _ _ _ _ _Unpleasant (measuring Appeal)
Stylish _ _ _ _ _ Tacky (measuring Identity)
Dull _ _ _ _ _ Captivating (measuring stimulation)

Seven semantic differentials are given for each dimension, making 28 differentials presented in random order and polarity. After the test, scores are summed up, and the average is calculated for each dimension. Several statistical tests have been performed on datasets from this questionnaire, and the researchers have found that the scales measuring ergonomic quality and hedonic quality are distinct, and that both contribute to the appeal (Hassenzahl, 2001).

A questionnaire like AttrakDiff is easy to administer to evaluators, is quickly done, and the designers may get feedback on whether the artefact is usable, interesting, and given a style that the evaluators feel comfortable with. These scores mean little in isolation, however. Averages towards the extremes are of course speaking a clear message, but averages towards the middle give little information about what worked. This is even more so when we consider that Likert scales are known to have a strong bias towards the centre. And even if the scores average towards "boring" rather than "interesting" (another example from AttrakDiff), there is no way of knowing what it is that makes the product boring, and whether different respondents were bored by the same aspects of the artefact.

We should also consider what a survey instrument like AttrakDiff actually measures. Hassenzahl's (2001) statistical analysis has shown that it is credible that the measures of identity, stimulation and ergonomics are separate, and that respondents appear to interpret them in consistent ways. However, we have not found that the authors have analyzed whether the semantic differentials actually capture these qualities. To be specific: When users state whether they find a product inventive, creative, bold, captivating, challenging and novel (the positive poles of the seven differentials for "hedonic quality: stimulation"), is there a systematic connection between their answers and the stimulation they experienced?

The two hedonic qualities that are measured with AttrakDiff are identity and stimulation, which are drawn from psychological literature. 'Stimulation' is the experience of a product to be new, enticing and challenging, while 'identity' is the feeling that owning a product would be an expression of themselves and a statement of which social group they belong to. These are believed by the authors to be of major importance when we experience a product as appealing or a joy to use, but could there not be other qualities that are equally important? Especially for genre design this is a pressing question, as Miller (1984) has shown that different genres are answers to different recurring rhetorical situations, that is, they serve different social purposes. Stimulation may be of importance for a pedagogical genre, but for other genres, we might equally well ask whether it inspired feelings of fun, tragedy or suspense — adjectives often used when describing genres in literature. We could apparently make questionnaires measuring fun, tragedy, and

suspense using semantic differentials, although it requires no little work to assure their validity as rigorously as Hassenzahl and colleagues have tested AttrakDiff. This work is justified for Hassenzahl as he believes the qualities they measure are universal, rooted in human psychology, and thus applicable to any product. Whether we can find such universal qualities for genre design, or indeed if we should believe in the possibility of such universals, is an open question.

If surveys give little detail, we should realize that the best way of accessing how users experience the text or service we are developing is probably to talk to them. Asking evaluators what they thought of the service can be a valuable source of information, but it needs to be carefully monitored. Our experience is that evaluators soon begin to suggest improvements to the service (see also Nielsen & Loranger, 2006). These are interesting, and should be collected, but one needs to be careful. We are not always aware of what makes us act in different ways, and what people think they would do in a hypothetical situation with a hypothetical artefact does not necessarily match what they in fact would be doing. More reliable are their reactions to the artefact, both emotionally and intellectually, and this is what the interviewer should be asking for.

In this section, we have given an overview of observation, surveys and interviews, all established methods in design science. From this short treatment, we notice two common traits: First, what these methods do best is to spot failures, or what in design literature is known as "critical incidents:" It is a critical situation when users aren't able to use the product as intended, give up and have to be helped, but also when they feel frustrated, bored or even angry. These are important results both for commercial design and for research. A design that users fail to understand will not do well in a commercial market, and finding such critical incidents early makes it possible to alter the design in order to avoid such failures. (In Akrich and Latour's terms, the test is a trial, ruling that the audience will not follow the scripts inscribed in the technology, so a new script should be devised (Akrich 1992, Akrich and Latour 1992).) For science wishing to establish universal principles of design, a failure could be considered a falsified hypothesis, which is the basis for knowledge in the many disciplines using the hypothetic-deductive method derived from Karl Popper's falsifiability criterion (Popper 1935/2005 p. 18).

The second common trait is that these methods mainly are based on comparison, whether explicit or implicit. It is difficult for a respondent in a think-aloud study to suggest improvements without pointing to another, existing product. A statistical measure, whether it is response times or average scores in a survey, is only meaningful when compared to the performance of another artefact, whether earlier versions or competing products.

We turn now to the other principle we proposed for evaluation in genre design: Evaluating with methods similar to those used to build the theories of communication and genre we built our designs on.

TOWARDS HUMANIST EVALUATION OF COMPUTER SYSTEMS

Humanist research is interpretative, not observational. Its objects are symbolic and meaningful structures made by human beings, such as writings, music, and visual art, and we who study these structures look for the possible meanings and aesthetic effects they create in readers, listeners, or viewers. Some interpretations aim at finding the exact intention of the author, others look at the meanings that are likely to be found by the audience. Sciences aim to explain and predict natural phenomena by principles that are constant; hence the metaphor of "laws" of nature. Texts and authors, on the other hand, are interpreted. Furthermore, each text or each work of art is unique, and cannot be explained by a general law (Gadamer, 2004). In the words of Wilhem Dilthey: texts are not explained, they are understood (Dilthey, 1883/1976 p. 89).

The test methods we know from systems design, human-computer interaction, and user experience design all assume a divide between system actions and user interface. All computer systems may be described in this way, and within digital media such as the Web or mobile apps, designers routinely describe this divide as "interface" and "content." A simple example of this divide may be a banking system, where the process of moving money from one account to another is sorted out without user involvement. From the bank customer's perspective, it just needs "to work," and the rules by which we judge whether it is "working" or not are known to everyone. Accessing your funds to pay bills is a matter of moving data in a database, although by no means a trivial system to build. Web media, such as a news site, a web TV channel or an online textbook are often also viewed as a problem of creating access to a database. Each text is viewed as principally similar, and an interface is made to find the text you wish for. Any text in the system will contain information, but only some text has the specific information that the reader wants. Reading is again turned into an access problem: a question of locating an answer within the database of texts.

Genre design is different. We are not designing access to generic "content"; we are creating new "content" that is significantly different from earlier "content," and it is this difference that we want to evaluate. There may be initial problems in handling the text with the interface control provided, and these problems may be addressed with evaluation methods from human-computer interaction. But when readers actually get to read the text, how do we evaluate the style, information, humour, drama, or pace — all the qualities that we appreciate when we study genres?

We should remember that texts always have been evaluated, but usually by a small group of experts. In publishing, experienced editors read novels, and coach their authors into making these novels better according to the editors' judgement. There is also a different issue in publishing: most publishers receive far more manuscripts than they are willing to publish. As such, the editor can already choose from a large number of prototypes. A similar abundance of manuscripts is found in the film industry, where only a tiny fraction ever gets filmed. In computer science terms, it is an iterative process: A manuscript is selected, re-written, a storyboard is created, before the film is shot, edited and edited again. In all phases there are evaluations in the form of readings and test runs. A first edit of the film may be showed to a test audience, and their reactions are used to judge whether the edit "works" or not. These are the kinds of evaluation methods we need to develop and make rigorous if we want to advance genre design as an academic practice.

So how can we perform a humanist evaluation? In our project, we have used three methods: Se-

mantic analysis of user interviews, within-subject A/B testing, and peer review.

User interviews

We have argued that user interviews are the most valuable evaluation method for genre innovation research, and most research projects in the literature have interviewed evaluators after an evaluation session. Krippendorff (2006, p.234) has suggested a more elaborate method for validation interviews: Early in the design process, we may ask stakeholders what characteristics a successful product or artefact should have. When evaluating the finished artefact, we ask stakeholders to describe it, and compare their descriptions with their earlier accounts of a desirable outcome. Similar descriptions indicate success.

Krippendorff inspired our evaluation of the Rome project. We interviewed evaluators after the test and asked them to describe the service in their own words before asking specific questions about the service. These descriptions were later analysed semantically to see if they matched our stated research goals and helped us in answering our research questions. We looked for descriptions and metaphors that gave insights into how the evaluators understood our service, or, as Krippendorff might put it, what our service meant to them. It has to be admitted, though, that the interviews did not yield a lot of insight. Our evaluators did not feel inclined to talk a lot of their impressions and meaning-making of the service, and it may be that Krippendorff is a bit too optimistic as to how verbal evaluators usually are. When asked to describe the service, they repeatedly described it as 'evocative' and 'informative,' and were at a loss to find more specific adjectives.

Another possible and related approach would be what Hartson & Pyla (2012) have called 'co-discovery,' where users evaluate a product in pairs, and the test is created in such a way that the evaluators have to talk to each other. Their conversations are recorded, and can later be analysed in much the same way as our interview data. This method is difficult to use with an application designed for individuals listening with headphones, however.

One may question whether it is possible at all to properly evaluate the finer details of text production in an interview. In our Rome project we have strived to find the right tone of voice, a suitable reading speed, how to relate historical information about the music with descriptions of its structure and tonality, and how much historical detail we should provide about each church. We have also discussed whether we should point out certain details within each church, or if we should limit the voice-over to descriptions that apply to all parts of the church. While we can find some comfort in the fact that all our evaluators reported that they liked our service, we are not likely to ever find a correct and verifiable answer to these deliberations.

Within-subject A/B testing

Advice and rules for readers are found throughout European history ever since the development of ancient Greek rhetoric, and it is always based on comparison. Some speeches, tragedies, letters, books, operas or films were clearly better than others, and scholars have studied them in detail to understand what made them so good. In a second iteration of our project, we introduced such a comparative aspect to our project. If the principles for locative audio we have deduced are robust, audience members should recognize that texts were less good when the principles were not followed. With this in mind, we have authored what could be likened to 'null hypotheses,' opposite examples, texts that purposely did not follow our own guidelines. This is similar in spirit to controlled experiments, what in web design is often called split design testing or A/B-testing (for an overview, see Kohavi, Longbotham, Sommerfield, & Henne, 2009), but we did a small-scale, qualitative study, not statistical testing.

In early April 2013, these new texts were evalu-

Table 1

For the summative evaluation, we created new texts in pairs to test the strength of some of the guidelines. We tested music with similar mood as the church room versus contrasting mood, structural similarities in music and architecture versus no similarities, and aura effect versus no mentioning of aura. Evaluators were asked to enter both churches in a pair, and explain which church they liked best and why.

Technique	Example	Counter-example
Music and commentary edited together, radio-like	Corelli's music in San Lorenzo in Damaso	Puccini's *Tosca* in Sant'Andrea della Valle Benevolo's and Cavallieri's music in San Luigi dei Francesi
Bringing music back to the original place	Corelli's music in San Lorenzo in Damaso	Benevolo's and Cavallieri's music in San Luigi dei Francesi
Music and church from same epoch	Corelli's music in San Lorenzo in Damaso	Puccini's *Tosca* in Sant'Andrea della Valle

ated in Rome. Eight volunteer students, none with musical training, were asked to visit two churches and listen to the program, constituting what we could call a "within-subject A/B test," as the text in one of the churches followed a principle, but not the other. Table 1 shows how the test pairs were constructed. After the test, evaluators were interviewed in a focus group.

This kind of test was able to bring out many nuances that the surveys and interviews did not capture. Our audio programs were edited in a form borrowed from radio: A piece of music with a catchy opening starts to play, and then fades to a lower volume as a narrator's voice is heard on top of the music. After a couple of minutes, the narrator finishes, and the music fades back up to the original volume. In the evaluation, we tried out a simpler approach: with music and commentary in different tracks. All evaluators strongly preferred the original combined version, one of them even created it himself, by starting the two audio tracks simultaneously. Weaker support was found for our initial assumption: That we should present music that was written for each church. We compared a program of Corelli's *Christmas concerto* in San Lorenzo in Damaso where it was first performed to a selection of Roman Baroque music in San Luigi dei Francesi, none of which had any historical ties to the place.

The evaluators did prefer the Corelli program, but did not state explicitly that it was because the music had been played there. Instead they commented on the narration, how it told a story about a person, and how it addressed the listener directly in the second person form. We still interpret this as support for music with historical ties to the place, if only because it lends itself to the kind of storytelling the evaluators appreciated. This could also serve as an example of the point we made above: That informants aren't always able to state clearly how they react to a prototype.

A third assumption of ours was that the experience would be most absorbing if the music was from the same period as most of the art in each church. Music history in Europe is parallel with the history of other arts, observing many of the same epochs and ideas, such as, e.g., Romanticism or Impressionism. Students of music history know this, while those with little or no training in classical music generally fdo not. Still, we hypothesized that a general audience would be able to appreciate a shared structure in music and architecture, a synesthetic appeal that did not require training. All the programs tested in the first round were constructed like this, and evaluators reported that they liked it. In our second round, we challenged this principle in San'Andrea della Valle, a church constructed from 1595 to 1650. The first act of Puccini's 1900 opera Tosca is set in this church, and we played three arias from Tosca, expecting the evaluators to find it literally out of place. They did not: They appreciated this selection just as much as the other programs they tested, clearly rejecting our hypothesis. This was the most striking result of our test, yielding new knowledge we could not have obtained without it. Thus, in this way we succeeded in replicating the approach of the natural sciences by being able to falsify an important hypothesis. While the hypothesis was based in the humanities, we were able to follow Popper's principle as originally expressed for the sciences.

Our experience indicates that this kind of comparison can be a valuable tool for evaluating text production research. Comparative texts can be tailored to answer the researcher's research question, and lead to more informative results than a statement that "evaluators liked what we hoped they would like". However, it requires a set of clearly operationalized research questions and a skilled interviewer to bring out the finer nuances of the evaluators' experiences.

Peer review

It is implicit in what we have written above that while an average user of a location-based system may feel very clearly what works or not for her/him, s/he may not be able to be very specific as to why it is so. Most audience members are not authors or analysts, and may never have given the finer details of locative writing much attention. What about other authors and critics? Peer review is a long tradition in scholarship. Just as scholarly articles are judged by a selection of peer reviewers, editors review books before they are published. A feature film regularly goes through stages of review and revision before its theatre release, both in the form of manuscript and early edits ("rough cuts") of the footage. To let other scholars analyse our texts (or services) should thus be a fruitful evaluation method, as Anders Sundnes Løvlie has argued and demonstrated earlier (2009).

To incorporate this aspect to our production, we performed an additional evaluation with a scholar with long experience in locative mobile media. His reactions did not match those of the student evaluators. Rather than describing the service, he suggested a long list of additions to it, stating that he wanted more of everything: More information, more music. He also suggested to split music and commentary into different tracks, as he only tested the combined versions. Our student evaluators explicitly took the opposite stand on all these suggestions: They preferred music and narration com-

bined, and they believed the length of commentary was suitable as they probably would not want to spend more time in the church.

Again, this was a revealing experience. At first, we should note that the expert reviewer did not comment on any problems with the service. Had he done so, it would have been our first priority to correct these. (This is known as an "expert inspection" in user experience literature, see, e.g., Hartson & Pyla, 2012). We must also realise that the expert reviewer's ideas are at least as good as ours, so to do proper research we should treat his suggestions in the same way and test them. But this also brings out the core problem for this kind of research: People are different and thus their tastes differ. Even if all scholars agree that Shakespeare is one of the greatest playwrights in history, many people do not care enough about his works to read them or see them performed. Our expert's tastes and expectations were different from the other evaluators, and this is to be expected when a target audience is diverse, as it is in this case. This experience indicates that peer review is an evaluation method that can provide rich and interesting data, but also that it is most valuable in early, formative evaluations, providing many alternatives to the existing design.

CONCLUSION

A research agenda for inventing new genres in order to gain new knowledge will need to use existing knowledge of texts, images, and sound for communication as its foundation. If we want to do design as research, we must take seriously Hevner, Park, March, and Ram's (2004) argument that research should aim to achieve new knowledge, to add something new to what we knew before. It will not suffice to do what is sometimes seen – namely, to build a system and write a paper that describes it, perhaps linking it to some earlier theory. That the author is pleased with his or her system hardly helps our common knowledge increase. We need to check our claims, put them to a test, and consider alternative explanations.

The established methods in design science deal with system functions and user interfaces, and are useful in the early phases of most design. But genres are about symbolic structures that create expectations and meaning in an audience; hence, we need to inspect such meaning-making properties when undertaking our research. In this paper we have proposed three such methods: Qualitative interviews that focus on the evaluators' meaning-making processes, systematic textual "experiments," painstakingly comparing one aspect after another, and peer review.

Human-computer interaction as a field belongs to a cognitive psychology tradition, and focuses on basic human perception, arguing that some laws of cognition are valid for all people. From this starting point, it is justifiable to do research with relatively small samples and conclude with universal guidelines. When we evaluate genre experiments, on the other hand, we soon realize that there are no universals in meaning-making.

People have different tastes, so there cannot be one single best solution for everyone. Large-scale quantitative studies is the only way to capture the variations in tastes with any reliability, but as we have argued above, these quantitative studies hardly give any insight into how and why one solution is preferred to another. A logical and ideal next step is thus to introduce an iterative research project, where qualitative inquiries into preferences and experiences are evaluated in randomized, large-scale A/B tests. I believe the present study has demonstrated how important this can be both for genre development and for theory building. This is a slow and costly process, however, and most genre experiments create small-scale prototypes, that can't be evaluated with hundreds of persons. More research is needed to find effective solutions to this *aporia*.

The digitalisation of media has made media studies approach computer science for concepts and understanding, but also increasingly for research methods. Digitial design is also a busy research area whith its own methods and approaches. I feel certain that media studies will adopt these approaches, and make design of media texts a common research strategy (cf. Fagerjord, 2012 and Nyre, 2014 for a more thorough discussion of this). It is my hope that these methods and other and better methods that will surface in the future can help those involved to do better research.

ACKNOWLEDGEMENTS

I am grateful for the many insightful comments throughout this project from Gunnar Liestøl, Anders Sundnes Løvlie, Carolyn Miller, Lars Nyre, the anonymous reviewers, and editors Charles Ess and Anders Olof Larsson. The department of Media and Information Science, University of Bergen kindly provided me with an office space where much of this text was written.

REFERENCES

Akrich, M. (1991). The De-scription of Technical Objects. In W. E. Bijker and J. Law (Eds.), *Shaping Technology, Building Society: Studies in Sociotechnical change*. Cambridge, Mass.: MIT Press.

Altman, R. (1984). Altman, Rick. "A Semantic/Syntactic Theory of Genre." *Cinema Journal 23* (3), 6-18.

Andersson, B.-E., & Nilsson, S.-G. (1964). Studies in the reliability and validity of the critical incident technique. *Journal of Applied Psychology*, *48*(6), 398. doi:10.1037/h0042025

Bolter, J. D. (2003). Theory and Practice in New Media Studies. In G. Liestøl, A. Morrison, & T. Rasmussen (Eds.), *Digital Media Revisited* (pp. 15-33). Cambridge, Mass.: MIT Press.

Dilthey, W. (1883/1976). "An Introduction to the Human Studies." In Rickman, H.P. (ed.), *Selected Writings*. Cambridge: Cambridge University Press.

Fagerjord, A. (2012). Design som medievitenskapelig metode [Design as method in media studies]. *Norsk medietidsskrift 19*(3). 198-215.

Fagerjord, A. (2011). Between Place and Interface: Designing Situated Sound for the iPhone. *Computers and Composition 28,* 255-63.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*(4). doi:10.1037/h0061470

Gadamer, H.-G. (2004). *Truth and method* (W. Glen-Doepel, J. Weisheimer, & D. G. Marshall, Trans. Second, revised ed.). London: Continuum.

Hartson, R., & Pyla, P. S. (2012). *The UX book: Process and guidelines for ensuring a quality user experience*. Amsterdam: Morgan Kaufmann.

Hassenzahl. (2000). Hedonic and Ergonomic aspects determine a software's appeal. *CHI Letters*, *2*(1).

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction*, *13*(4), 481-499.

Hevner, A. R., March, S. R., Park, J., & Ram, S. (2004). Design science in information systems work. *MIS Quarterly*, *26*(1), 75-105.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, *18*(1), 140-181. doi:10.1007/s10618-008-0114-1

Krippendorff, K. (2006). *The semantic turn: A new foundation for design*. Boca Raton: Taylor & Francis.

Krug, S. (2010). *Rocket surgery made easy: The do-it-yourself guide to finding and fixing usability problems* (Kindle ed.). Berkeley, California: New Riders.

Latour, B., & Woolgar, S. (1986). *Laboratory work: The construction of scientific facts* (Second ed.). Princeton University Press.

Lewis, C. (1982). *Using the 'thinking-aloud' method in cognitive interface design*. Research report RC 9265. IBM TJ Watson Research Center.

Liestøl, G. (1999). Rhetorics of Hypermedia Design. In *Essays in Rhetorics of Hypermedia Design* (Ph.D. Dissertation ed., p. 265). Oslo: Department of Media and Communication, University of Oslo.

Liestøl, G. (2006). Conducting Genre Convergence For Learning. *Cont. Engineering Education and Lifelong Learning*, *16*(3/4), 255-270.

Liestølg, G. (2009). Augmented Reality and Digital Genre Design: Situated Simulations on the iPhone. R. Gasset, C. Disalvo, J. Pari and J. Bolter (Eds), *8th IEEE International Symposium on Mixed and Augmented Reality: Arts, Media and Humanities Proceedings*. doi: 10.1109/ISMAR-AMH.2009.5336730

Løvlie, A. (2009). Textopia: Designing a locative literary reader. *Journal of Location Based Services*, *3*(4), 249-276.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*, 251-266.

Miller, C. (1984). Genre as Social Action. *Quarterly Journal of Speech 70,* 151-167.

Moggridge, B. (2007). *Designing Interactions*. Cambridge, Mass: MIT Press.

Moulthrop, S. (2005). *After the last generation: Rethinking scholarship in the age of serious play*. Proceedings from Digital arts and culture, Copenhagen.

Neale, S. (2006). *Genre*. London: British Film Institute.

Nelson, T. H. (1974). *Computer Lib/Dream Machines*. Self-published.

Nelson, T. H. (1992). *Literary Machines 93.1*. Sausalito: Mindful Press.

Nielsen, J. (2000). *Designing Web Usability: The Practice of Simplicity*. Indianapolis: New Rider.

Nielsen, J., & Loranger, H. (2006). *Prioritizing Web Usability*. Berkeley, California: New Rider.

Nyre, L. (2014). Medium design method: Combining media studies with design science to make new media. *The Journal of Media Innovation*, *1*(1), 86-109.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Popper, K (1935/2005). *The Logic of Scientific Discovery*. London: Routledge.

Ulmer, G. (1989). *Teletheory: Grammatology in the Age of Video*. New York: Routledge.

APPENDIX

Survey questions used in the evaluation of the Musica Romana (translated from Norwegian)

1. How did you like this service?

Not at all - not very much - neither much or little - quite a bit - very much

The respondents were asked to score how much they agreed to the following statements on a five-point Likert scale ranging from "totally disagree" to "agree totally".

2. "The talking disturbed me. It would be better to just have the music."

3. "The music made me experience the church in a different way."

4. "The comments from the narrator made me experience the music different from how I otherwise would have."

5. "The music does not fit with the church the way it looks today".

6. "I feel I understand the history of the church now".

7. "When the narrator spoke, I often 'was lost' and though about something else."

8. "I learned something about music history".

9. "It was exciting to be in a place where the music was performed originally".

10. "The whole thing was boring"

11. "I would like to try this service in more of Rome's churches."