“Clementine will remember that” – Methods to Establish Design Conventions for Video Game Narrative

Hartmut Koenitz
HKU University of the Arts
Postbox 1520
3500 BM Utrecht, The Netherlands
+31 (30) 209 1509
Hartmut.koenitz@hku.nl

Christian Roth, Noam Knoller, Teun Dubbelman
HKU University of the Arts
Postbox 1520
3500 BM Utrecht, The Netherlands
+31 (30) 209 1509
{Christian.roth,noam.knoller,teun.dubbelman}@hku.nl

ABSTRACT
In this paper, we describe narrative game design as an area for empirical research and aim to promote additional work in this area. The focus of our paper is therefore on the process. We start by discussing the relationship between the design of the narrative aspects of video games vs. non-narrative aspects, as well as in comparison to earlier narrative media. On this basis, we identify specific challenges from the perspective of design. Then, we define “design conventions” and introduce our method for identification and verification using empirical methods. In this context, we discuss methodological issues and advocate best practices. Finally, we report on early results and outline future work.

Keywords
video game design, narrative design, ludonarrative, design conventions, empirical methods

INTRODUCTION
“Clementine will remember that” – this notification in The Walking Dead Game (Telltale Games 2012) informing the player that her actions will have consequences much later on, has become a calling card for the “Telltale Formula” – a collection of design conventions that have made the company’s narrative-focused games a success with critics and audiences. But what are narrative design conventions, actually? Do they differ from other design methods in game design, supposedly the ones that are focused on rules? Does such a differentiation make sense from the design perspective? Or should we rather understand some design methods to be usable in both ways, as Dubbelman (2016) can be understood to have argued for? Certainly, there is a difference of attention – a simple search for professional books on narrative design in comparison to general game design shows a clear emphasis on the latter.[1] In addition, it is safe to say that at least some game designers

Proceedings of DiGRA 2018

© 2018 Authors & Digital Games Research Association DiGRA. Personal and educational classroom use of this paper is allowed, commercial use requires specific permission from the author.

understand narrative in the sense of its traditional immutable manifestations like the novel or the movie, and thus as a challenge for dynamic, rule-based game design. For example, Bernd Kreimeier distinguishes games from earlier "narrative media":

> [...] designers have worked around this deficiency by relying on techniques and tools borrowed from other, older media – [...] narrative media like cinematography, scriptwriting and storytelling. [...] However, the metaphors and devices borrowed from narrative media are usually insufficient (or even inadequate) to capture the essence of the interactive game medium (Kreimeier 2002).

This perspective from 2002 might contain echoes of the then contemporary narratology vs. ludology debate (which started with a rejection of games as narrative media). However, when it comes to narrative, even in 2017, game designers fall back on design methods established in earlier media. For example, in an interview for Gamasutra, game designer Tariq Mukhttar references the "8 Point Story Arc methodology" – from Nigel Watt's book on how to write a novel – as a major influence on his level design:

> Midway through development I employed the '8 Point Story Arc' methodology to test how well the narrative plays out. It forced me to make some big changes to the level design (LeRay 2017).

While this approach – leveraging long-established design methods for video game design – might seem to make sense from a pragmatic perspective, it also creates a problem by making game narrative a derivative form. And here lies danger: as long as narrative game design relies on the methods of the novel or the movie, it will invite unfavorable comparisons to the original. Indeed, as Ian Bogost reminds us, "Video games are better without stories. Film, television, and literature all tell them better" (Bogost 2017). Should we be content with the derivative and diminished role of narrative in video games? The answer to that challenge, as Koenitz has argued earlier (2016), lies in a more inclusive understanding of narrative, a realization that film, television and literature only represent a small part of the overall space of narrative manifestations and that there is plenty of room for genuine interactive digital narrative expressions, of which game narrative is one variety. To rephrase Bogost: 'Video games are better without the kind of stories that are native to film, television, and literature. Those media will always tell them better'.

THE CHALLENGE OF NARRATIVE GAME DESIGN

Once we no longer understand narrative game design as an attempt to achieve the same effects that attract us to film, television, and literature, two things happen simultaneously: first, narrative games would no longer have to compete "on the other guys' turf" as their narrative aspects would no longer be compared directly to literary or cinematic narrative and would instead vie for audiences on the basis of their own particular strengths as a separate narrative medium. Second, we enter an exciting, but much less defined space of specific narrative game design in which we can no longer automatically assume the applicability of established practices from earlier narrative media.

Creation of such interactive narrative experiences is challenging, as new design methods have to be invented and successfully implemented. Conversely, Janet Murray regards the "invention and refinement" of design conventions as a focus area for research in digital media as an expressive practice (Murray 2012). What has been missing so far is an effort to formally identify and share a broader range of emerging conventions and overarching design principles specific to narrative games. In addition, there is the danger of

"unproductive attempts to apply legacy conventions to new digital frameworks" (Murray 2012), which fail to exploit the expressive potential of interactive media and thus result in unsatisfying products. Consequently, a high obstacle exists for newcomers who enter the field, as there are barely any design guidelines to learn from. In addition, little formal training in the narrative aspects of games exists. Professionals in the role of Narrative Designers are mostly self-trained, as many Game Design programs offer a single course on narrative aspects in a typical multi-year program. As of early 2018, only a handful of programs offer a degree in "game writing" or similar specialization.

To better understand this space of interactive narrative design, we start by considering the narrative aspects of the player experience. Within narrative games, players take action and make meaningful decisions, thus altering narrative progression and outcome. This description might not sound particularly different compared to games not focused on narrative. However, the discussions around games like *Dear Esther* (Pinchbeck 2008), *Gone Home* (The Fullbright Company 2013), and *Firewatch* (Campo Santo 2016) show that many players and commentators do make a distinction. For some, such games are annoying "walking simulators", devoid of the action-related pleasures they expect from something carrying the 'game' label. For others, however, these games' particular attraction lies precisely in their focus on narrative, in the way they afford specific interactive experiences. It is certainly difficult to define the difference precisely in terms of design, yet some tendencies are apparent: instead of winning or losing, of overcoming a particular challenge to survive a level, the focus is on decision-making for the purpose of character development, to uncover secrets and unknown parts of a prior narrative, and to propel a developing narrative forwards. From a more abstract point of view, we can therefore describe this kind of interactive narrative design as being concerned more with grey areas and overall trajectories than with the design of first-person shooter games or a platformer, where players either survive a level or not and where the design affords a binary win/lose situation. From a more concrete perspective, the distinction is much less clear-cut. *Firewatch*, for example, features combination locks for the player to open, steep inclines to climb and fences to scale – features that would not be out of place in games devoid of narrative focus. However, in *Firewatch*, all of these features are in the service of the overall narrative experience and never feature prominently on their own. A way to identify narrative design might therefore be to pose the question: 'does an aspect of the design stand alone and is perceived so by players, or does it serve the overall progress of character development and/or narrative'?

LEVELS OF ANALYSIS: DESIGN PATTERNS, DESIGN PRINCIPLES, AND DESIGN CONVENTIONS.

The concept of "patterns," derived from architecture, features prominently in video game design research (Barwood and Falstein 2002; Björk and Holopainen 2004) "semi-formal interdependent descriptions of commonly reoccurring parts of the design of a game" (Björk, Lundgren, and Holopainen 2003) as Björk/Holopainen put it. Unfortunately, considerable differences exist in the respective definition of 'patterns,' as Kreimeier reminds us (Kreimeier 2002). A significant drawback of game design patterns is therefore the lack of a precise (and shared) definition that would allow for direct comparison between different collections. For example, the description of the Rock-Paper-Scissors pattern differs considerably between Kreimeier and Björk/Holopainen, which becomes evident already in the abridged form reproduced here for comparison. First Björk/Holopainen, then Kreimeier:

PAPER ROCK SCISSORS

Description: This pattern is based on the children's game with the same name. It means that players try to outwit each other by guessing what the other ones will do, and by tricking other players to take a wrong guess on one's own action. The original game is very simple; after a count to three both players make one out of three gestures, depicting rock, paper or scissors. Rock beats scissors, scissors beat paper and paper beats rock. That there is no winning strategy is the essence of the pattern: players have to somehow figure out what choice is the best at each moment.

This game pattern is well-known with the game design community (sometimes called "triangularity", see Crawford) and is a mnemonic name for the logical concept of non- transitivity (basically, even if A beats B and B beats C, A doesn't beat C).

Examples: Quake (relation between weapons and monsters), Drakborgen, SimWar, protogame to show non-transitivity (Dynamics for Designers, Will Wright, GDC 2003) […] (Björk and Holopainen 2004)

PAPER-ROCK-SCISSORS
Problem: Avoid a dominant strategy that makes player decisions a trivial choice.

Solution: Introduce nontransitive relationships within a set of alternatives, as in the game of paper-rock- scissors.

Consequence: The player is no longer able to find a single strategy that will be optimal in all situations and under all circumstances. She has to revisit her decisions, and, depending on the constraints imposed by the game, adjust to changing situations, or suffer the consequences of an earlier decision. […] (Kreimeier 2002).

Kreimeier's patterns have the descriptors Name, Problem, Solution, Consequences, Example, Björk/Holopainen instead use Name, Description, Consequences, Using the Pattern and Relations. While there is certainly some overlap between these different categories, one cannot simply be used to extend the other and a comparison becomes challenging. Another angle of criticism towards this approach stems from its combination of a high level of abstraction with concrete examples. What is missing between these two aspects is concrete, but transferable, design knowledge. Conversely, Koenitz in 2015 diagnoses a void between abstract descriptions and particular examples when it comes to interactive narrative design:

> [a] high level of abstraction provides little concrete design advice. [In contrast], the highly specific nature of particular projects often make it difficult to identify generalizable conventions (Koenitz 2015).

Seen from this perspective, design patterns – at least in the way they are currently conceptualized – are unable to fill this void. Instead, we propose two levels of analytical categories: abstract 'design principles' and concrete 'design conventions.' An example for a design principle is Murray's "scripting the interactor" (Murray 1997) by which the interactor is made aware of the overall context and interactive potential of her role. Existing

on a higher level of abstraction, design principles invite a range of different implementations. In the case of "scripting the interactor", a concrete implementation could be the use of textual introductions at the beginning of the experience. This constitutes the intermediate level of transferable design knowledge, while the concrete appearance and words used in a given example represent its particular instantiations. We call this kind of transferable knowledge a 'design convention': a design method that creates a particular effect (a conventional understanding) in the audience. The remainder of this paper describes our approach on how to establish Design Conventions by means of empirical methods; more exactly the identification and experimental verification of their effects on the user experience.

Before we can finish this section, we would like to acknowledge that our perspective on design conventions has some conceptual overlap with "Strong Concepts" in HCI:

> Strong concepts are design elements abstracted beyond particular instances which have the potential to be appropriated by designers and researchers to extend their repertoires and enable new particular instantiations (Höök and Löwgren 2012).

However, Strong Concepts differ in that they focus on the individual designer's mastery in their application and their perspective on audience reaction, which is taken as more varied and unpredictable.

IDENTIFYING AND EVALUATING NARRATIVE GAME DESIGN PRINCIPLES AND DESIGN CONVENTIONS

In this section, we first consider scholarly approaches towards the identification and evaluation of narrative game design principles and design conventions. Then, we describe best practices, explain our concrete setup and report on early results. The purpose of our approach is to increase the body of accessible and transferable knowledge.

Empirical methods provide means to gather knowledge based on actual artefacts, as well as players' reactions to them. Both qualitative and quantitative methods have a place in interactive narrative design research and should be combined for maximum effect. We recommend mixed-method approaches for a holistic understanding of the effects of design methods on the user experience, combining qualitative and quantitative research, using explicit, subjective (interviews, questionnaires) and implicit, objective data (physiological measurements, statistics from artefacts). Of equal importance is a clear understanding of what aspect of design is to be identified and verified. As we have argued in the preceding section, "design patterns" is too abstract a concept to allow concrete measurements. Instead, we focus on more fine-grained and concrete design methods, which we call "design conventions" in contrast to more abstract "design principles". Our approach is divided into four broad phases: identification of candidates, selection of method, user study setup and execution, interpretation and verification of results.

Phase 1: Identify design principles and convention candidates

Qualitative research provides an important starting point by identifying abstract design approaches and concrete implementations, that can be verified in a next step within, quantitative studies on larger samples. Qualitative methods (content and design analysis) can, for example, identify design convention candidates in critically acclaimed narrative games. Phenomenological approaches as well as auto-ethnographic methods (self-reflection and writing) deliver starting points for further investigation. Additional material is available in professional game design literature and public talks (e.g. at the GDC

Narrative Summit) and post mortem presentations. Furthermore, focus group interviews with narrative game designers, e.g. by means of semi-structured questionnaires, will result in a collection of subjective design methods. These initial approaches do not require a lot of participants to be insightful. Focus group interviews can be conducted with different target groups, with group sizes as small as four. Analyzing user reviews of narrative games (e.g. on Steam, metacritic) can point out design successes and flaws. The resulting collection of design methods will then be analyzed for commonalities to identify design convention candidates.

It is important to be aware of the danger of 'false positives' in this phase. Frequency alone is not a sufficient criterion to declare a certain design method a convention candidate. Conceptually, our understanding must have room for "design fads" – frequently applied, but actually ineffective design methods. In addition, we should be aware that literary and cinematic design conventions, transferred to narrative games, are essentially foreign objects in a very different context. This means we should be careful when drawing similarities to conventions from non-interactive media and carefully scrutinize their actual application and impact under interactive conditions. In the next step, quantitative methods are then used to verify the effectiveness of these candidates.

Phase 2: Methods to evaluate convention candidates

Psychological perspectives are a major factor in evaluating current and future interactive digital narrative systems. The appreciation of current and future narrative games and their commercial success will be related to the purposeful application of design conventions to create satisfying and fulfilling experiences. This means evoking satisfaction or frustration by meeting, manipulating and subverting target audience expectations. User expectations do not merely precede such systems but are co-created in an iterative process of system design, experience, evaluation and feedback. Authorial considerations may justify the frustration and subversion of such expectations, leading once again to a readjustment of such expectations.

For the evaluation, two main experimental setups exist: within-subject and in-between-subject. The within-subject setup exposes participants to different test conditions in sequence. To avoid possible sequential effects, the order of the conditions should be varied. In an in-between setup, participants are divided into subgroups, and each group experiences a different test condition. Data representing the user experience is usually acquired by means of questionnaires, which feature validated scales measuring different experience dimensions, such as flow, presence, and enjoyment. The advantage of this study design is that it can be easily administered. Also, asking participants directly about their experience has high reliability and validity. Furthermore, administering a questionnaire after exposure is non-intrusive as it does not interrupt the participant's experience.

However, disadvantages of post-hoc measurements also exist, chiefly the lack of information on temporal variations of the user experience. Yet, these are relevant for research in interactive narrative design, because good narratives, by their very nature, feature different pacing and thus elicit a range of affective responses over time. Indeed, regarding video game design, Pagulayan, et al. (Pagulayan et al. 2009) remind us that the success of a play environment is determined by the process of playing, not its outcome. Post hoc questionnaires can ask participants to assess their experiences during gameplay, yet these experiences might be hard to recall in a precise manner. This is especially problematic for experiences lasting longer than half an hour. In addition, participants might go through phases of different experiences during exposure. For designers, it is often

crucial to identify "unattractive" sequences. Schønau-Fog tries to address this issue by means of interrupting the game play for interactor feedback [46]. However, this kind of intrusive measurements can severely disrupt the experience (cf. [19]), which is especially problematic when the researcher wants to obtain data about flow and presence. This means that studies focusing on these temporal aspects represent an opportunity for future research. A promising approach is in 'diegetic measurements,' by which we mean a further hiding of the scientific aspect of the study and more seamless integration into the virtual world of the narrative game experience, e.g. by having players file a report from within the game. Naturally, the availability of this approach is highly dependent on the particular game. Physiological measurements during game experience are also a promising route, yet they pose additional problems by creating an abundance of data that can be difficult to interpret.

For our studies we use Roth's measurement toolbox, which addresses a range of relevant user experiences in the context of narrative games with validated, distinct scales in post-hoc questionnaires. This framework enables the measurement of narrative game user experience dimensions on a quantitative level. More recently, Roth's dimensions have been aligned and recast as components of Murray's experiential qualities of agency, immersion, and transformation (Roth and Koenitz 2016). By making Murray's categories 'measurable,' this connection has the potential to enhance the dialogue between more practice-oriented perspectives and the humanities inside game studies.

Phase 3: Study setup and execution

To experimentally verify the effectiveness of particular design convention candidates we devise a predominantly quantitative approach that connects the creation of prototypes with the evaluation of user experience. We apply this approach to test the effectiveness of design strategies by comparing prototypes that differ only in one specific characteristic (A/B). For example, in a first study, different varieties of text-based conditioning/scripting at the beginning of a game showed that exaggerated claims can significantly lower player identification with their game character compared to descriptions that were validated by the gameplay.

In our case, participants were recruited via email and randomly distributed to online experiment conditions, namely different versions of a narrative game. After interacting with this artefact for a variable time (usually 10 to 20 minutes), participants were guided to an online questionnaire, presenting statements that must be agreed or disagreed to, on a 5-point Likert scale (quantitative measure). Additionally, participants were asked to freely write about positive and negative aspects of their experience (qualitative measure).

Best practices for user studies

Empirical methods can give us insights into the effects of particular design decisions. For user studies to be effective, several best practices should be followed. First of all, we advocate to always start with a small pre-study to test the overall setup, to see how participants react and to get a first indication of possible results. This might seem like an extra step, yet, in fact, it is less costly in terms of effort and time to make adjustments at this stage. The next step is to carefully plan the actual study.

For the purpose of the following discussion, we take as an example a typical quantitative user study comparing prototypes that only differ in a particular narrative design approach. For instance, one prototype featuring diegetic notifications commenting on player agency and another prototype without this feature. In this example, the goal of the study would be to measure the effects of that particular design method on the user experience in terms of

perceived narrative agency, immersion (flow, presence) and enjoyment. Regarding sample size, the rule of thumb for a minimum of 25-30 participants per group would be applied. Therefore, at least 50 participants would be recruited and randomly assigned to the two test conditions, i.e. each prototype version would be play-tested by at least 25 participants. Both groups would then fill out an identical digital questionnaire containing user experience scales as well as open questions. In analyzing participant responses, the study should also take into consideration the possible need for data set exclusion. This is due to the potential discrepancy caused by the results, for example, of 4 participants (2 from each group) turning out to be invalid due to technical issues, lack of qualification, [2] etc., hence causing the final sample size to be a total of N = 46.

The problem with this example study lies in its small sample size of 46 participants and two test conditions. To understand why the study is problematic, we need to look at the relationship between sample size and effect size. Effect size is "a way of quantifying the size of the difference between two groups. […] It is particularly valuable for quantifying the effectiveness of a particular intervention, relative to some comparison. It allows us to move beyond the simplistic, 'Does it work or not?' to the far more sophisticated, 'How well does it work in a range of contexts?'" (Coe 2002). Cohen (Cohen 1992) suggests, when comparing mean values of two groups, to interpret effect sizes of .20 as small, .50 as medium, and .80 as large effects. It is important to note that even small effects can be substantial. For an in-depth discussion on the importance and interpretation of effect sizes, see the papers by Robert Coe (Coe 2002) or Sullivan and Feinn (Sullivan and Feinn 2012). Effect sizes are easy to interpret since they are the equivalent to a Z-score of a standard normal distribution; hence an effect size of .80 means that the score of the average person in the experimental group is .8 standard deviations above the average person in the control group (Coe 2002). In order to compare effect sizes between studies with different participant sizes, we can use a specific unit that exists for this purpose, *Cohen's d*.

This might seem all good in principle, but how can we apply this knowledge in the design of studies, as the relevant measurements only exists post hoc? Before setting up a user study, we can use G*Power (Faul et al. 2009), a software package that helps us gain the necessary understanding beforehand. More exactly, G*Power can calculate the power of a study, given a particular effect size and sample size. For example, we might assume a medium effect size of $d = .5$, measuring the impact of diegetic notifications on the player's perceived narrative agency. Setting the cut for the significance level at the common $p = .05$ (one-tailed), with 46 participants in total and a between-subjects study setup, a priori calculated power is only .51, which translates into a 49% chance of not finding a significant effect even when there is one (a so-called Type II error). This means that by just looking at significance, one in two studies with this setup would not find a medium effect. Similarly, Bakker, van Dijk, & Wicherts (Bakker, van Dijk, and Wicherts 2012) criticize the common behavior of using 25-30 participants per test condition as an underpowered study design. This is problematic for two reasons: first, small sample sizes increase the bias and the likelihood of inflated effects based on chance. Second, studies with small sample sizes lack the statistical power to find significant effects even though a genuine effect exists in the population. It is therefore crucial to understand that significance tests of p-value depend not only on the size of the effect but also on sample size. What makes this issue even more crucial is the fact that based on our experience, the actual effect sizes of different design methods are often small to middle sized, thus requiring much larger samples. Consequently, significant results could exist but when the effect is rather small, they are only detectable in adequately large sample (Cohen 1995). Since statistical significance does not by itself include information about the size of an effect, we encourage fellow scholars

to always report effect sizes (e.g. Cohen's measures[3]), which measure the strength of a result and do not depend on sample size. At least for the reporting of main effects, this should become common practice.

Phase 4: Interpretation and further validation

When it comes to results, the first question should be: do the results make sense? Can significant differences between the conditions and effect size of the difference be explained theoretically? Do they verify or nullify the initial hypothesis? If they do not – why? In the next section, we offer some potential explanations for discrepancies. In general, replications of studies are needed to support concrete findings regarding the effects of certain design choices on the user experience. As Stroebe and Strack (Stroebe and Strack 2014) remind us, exact replication of an experiment would operationalize both dependent and independent variables in exactly the same way as the original study. At the same time, variations of studies are needed, as the analysis of design convention candidates is crucial to test a specific narrative mechanic with different narrative game designs and different target audiences. So, based on context and implementation, we have to assume that the same narrative game mechanic may potentially have different effects on user experience.

Additional Factors impacting a study's validity

Lab studies are cost intensive and time consuming, whereas online studies allow for large sample sizes that can potentially include observation over time, using inexpensive built-in laptop or mobile device cameras recording participant emotions. A disadvantage of online studies is the lack of control over external factors, such as time of day of participation in the study, computer system and peripherals used, distractions from the surroundings, etc. This might limit the internal validity of the study. However, these mitigating factors are contrasted by several practical advantages: participants can decide for themselves when and where to play the game, much larger groups of participants can be reached, and cultural and regional backgrounds can be more diverse, resulting in higher external validity. It is also crucial to have comparable groups for test conditions, and to use random assignment of participants to experiment conditions, so that the only differences between groups would be due to chance. These so-called 'true experiments' can be used to investigate cause-and-effect relationships and provide high internal validity.

An important determining factor for users' experiences is their prior engagement and frequency of contact with video games. This pre-existing knowledge has direct impact on the experience through genre expectations and familiarity with controls, like the WASD keys. For example, a group of students working with us found the controls of *Firewatch* as being too difficult to figure out for non-gamers in the time allotted for the study and thus decided to instead choose *Life is Strange* (Dontnod 2015). At this point, we cannot yet exclude an influence of reported gendered roles and therefore, an unequal distribution in this regard between experiment groups might influence the test results. To overcome this issue in studies with in-between subject design and different experiment conditions, participants need to be randomly assigned to the conditions while monitoring the balance of gender role distribution. Before any further analysis, mean values of gender, age and computer literacy should be compared to guarantee no significant attribution differences between the groups.

When conducting user experience research, it is crucial to test with all target groups in mind. Samples should represent the population or a specific subgroup. However, user studies are often conducted with participants that are easily available, often students enrolled at the same university, instead of a more inclusive group needed for a

representative outcome. More concretely, the use, for example, of game design students as subjects might not create a valid representation of the population and thus can result in limited external validity.

Another factor that might limit the external validity of user experience studies is the use of student-produced game prototypes, which usually cannot compete with commercial games produced over the course of several years by large teams in professional game studios. Finally, play sessions are often rather short (5 to 20 minutes) and therefore might not represent typical narrative game experiences, which are designed to take hours to complete. However, study setups with very long test sessions come with their own limitations: participants might lose focus over time, and post-hoc measurements are less reliable as participants will only remember parts of the experience,

Early Results

So far, we have conducted a number of studies with our methodology. For our first study, we created an A/B/C setup, evaluating different variants of introductory text (neutral, precise description of player's agency and over-promising) meant to script interactors into their role. This online study had 60 participants, evenly divided amongst reported gender roles. We found players' perceptions were significantly affected by the different variants: scripting that over-promises in terms of interactor control diminishes identification with the player character while precise scripting leads to higher perceived agency. Two follow-up studies extended that study. The first follow-up pilot study evaluated the impact of the original interactive multimedia introduction of the narrative game *Firewatch* in comparison to two alternative text introductions, one basic, neutral introduction and one scripting players for narrative agency. Participants played the game for up to an hour before filling out the same questionnaire used in our original study. Results indicate that being scripted for narrative agency resulted in a significantly higher perception of autonomy. The second pilot study investigated the effects of introductions, which were phrased to evoke different emotional responses, on the level of engagement users had when playing the narrative game "Life is Strange" for about 40 minutes. This study introduced physiological measures of skin conductance and heart rate to our setup, but did not show any significant results yet, due to the limited sample size of only 17 participants in total.

CONCLUSION AND FUTURE WORK

In this paper, we have introduced our notion of narrative design conventions and outlined our approach using empirical methods to identify and verify them. One of our emphases has been on describing best practices for user studies, with the aim of promoting additional work. In addition, we hope to provide a reference point for improving user studies in games research. So far, sample sizes have often been too small for meaningful statistical analysis (Bakker, van Dijk, & Wicherts, 2014). Before conducting a study, a power analysis of the study design is needed, but often not carried out. Studies with low statistical power can lead to the false conclusion that there are no effects present, when in reality the sample size was not large enough to allow for the detection of effects (Type II error). Likewise, exceedingly large sample sizes can lead to an overpowered setup, showing significant results that are in fact not necessarily meaningful (Type I error). The free tool G*power can help to prevent these errors, by performing power analyses a priori, based on the expected effect size, e.g. when comparing the effect of one design method over another.

As researchers, we do understand the pressures and constraints under which many of us operate. In practice, this translates to compromises, also when it comes to user studies. However, games researchers should strive to the highest possible standard under the

circumstances. The minimal requirement in this regard would be to aim at transparency regarding the limitations that a particular study encounters. For example, a study that does not meet all the requirements could be presented as a "pilot study" with the idea of following up with additional steps (i.e. replication studies for verification) later. In particular, we strongly encourage the reporting of effect sizes as a standard, to help with comparison between studies.

With regards to narrative design conventions, much work remains to be done. We invite fellow researchers to join us in an effort to identify, evaluate and contribute to a body of design knowledge in this way. Beyond that, many additional questions remain. For example, do design conventions depend on specific contexts, e.g. do adventure games afford certain conventions that would be out of place in other narrative games? Another interesting area to explore is the overall scope of such conventions beyond games – e.g., whether they are valid for other forms such as interactive documentaries.

ACKNOWLEDGMENTS

We acknowledge the following Amsterdam University College Students' work on the follow-up studies mentioned in the paper: Sam Braun, Soraya Duncan, Pauline Hageman, Gordon Lucas, Lyanne van der Plank, Tessa Scholten, Vani Ahuja, Manon Blanke, Pieter Buis, Daniek Joosten, Christina Keßler, Zuzanna Orlowska.

ENDNOTES

1 A search on amazon.com (Feb 8, 2018) yields more than 1000 results for books on the exact term "Video Game Design." In contrast, no books are found for "Interactive narrative design," or "narrative game design." Only if search terms are relaxed ("game narrative" design) 12 books are found, and 10 for ("interactive narrative" design).

2 NB: Always test more participants than you aim for, as you might need to exclude a few datasets. Some participants might not take the testing seriously, which can be detected as patterns in their answering behavior in the questionnaire.

3 Different statistical tests require specific effect size measures. For the comparison of group mean values, Cohen's d became a frequently used measure to estimate effect sizes and required sample sizes.

BIBLIOGRAPHY

Bakker, Marjan, Annette van Dijk, and Jelte M Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7 (6): 543–54. doi:10.1177/1745691612459060.

Barwood, Hal, and Noah Falstein. 2002. "The 400 Project." http://www.theinspiracy.com/the-400-project.html.

Björk, Staffan, and Jussi Holopainen. 2004. *Patterns in Game Design (Game Development Series)*. Charles River Media.

Björk, Staffan, Sus Lundgren, and Jussi Holopainen. 2003. "Game Design Patterns." DIGRA, Level Up conference, Utrecht 2003

Bogost, Ian. 2017. "Video Games Are Better Without Stories." *Theatlantic.com*. April 24. https://www.theatlantic.com/technology/archive/2017/04/video-games-stories/524148/?utm_source=atlfb.

Campo Santo. 2016. "Firewatch [Video Game]." Portland, OR: Panic.

Coe, Robert. 2002. "It's the Effect Size, Stupid: What Effect Size Is and Why It Is Important," September. Exter, UK: Education-line. http://www.leeds.ac.uk/educol/documents/00002182.htm.

Cohen, Jacob. 1992. “A Power Primer..” *Psychological Bulletin* 112 (1): 155–59. doi:10.1037/0033-2909.112.1.155.
Cohen, Jacob. 1995. “The Earth Is Round (P < .05): Rejoinder..” *American Psychologist* 50 (12): 1103–3. doi:10.1037/0003-066X.50.12.1103.
Dontnod Entertainment, 2015. “Life Is Strange [Video Game].” Dontnod Entertainment.
Dubbelman, Teun. 2016. “Narrative Game Mechanics.” In *Interactive Storytelling*, edited by Frank Nack and Andrew S Gordon, 39–50. Springer International Publishing. doi:10.1007/978-3-319-48279-8_4.
Faul, Franz, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. “Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses.” *Behavior Research Methods* 41 (4): 1149–60. doi:10.3758/BRM.41.4.1149.
Höök, Kristina, and Jonas Löwgren. 2012. “Strong Concepts.” *ACM Trans. Comput.-Hum. Interact.* 19 (3): 1–18. doi:10.1145/2362364.2362371.
Koenitz, Hartmut. 2015. “Design Approaches for Interactive Digital Narrative.” In *Interactive Storytelling*, 9445:50–57. Lecture Notes in Computer Science. Cham: Springer International Publishing. doi:10.1007/978-3-319-27036-4_5.
Koenitz, Hartmut. (2016). Interactive Storytelling Paradigms and Representations: A Humanities-Based Perspective. In Handbook of Digital Games and Entertainment Technologies (pp. 1–15). Singapore: Springer Singapore. doi: 978-981-4560-52-8
Kreimeier, Bernd. 2002. “The Case for Game Design Patterns.” *Gamasutra.com*. March 13. http://www.gamasutra.com/features/20020313/kreimeier_01.htm.
LeRay, Lena. 2017. “Developing a Cat's Manor in Saudi Arabia for a Western Audience.” *Gamasutra.com*. January 3. https://www.gamasutra.com/view/news/285904/Developing_A_Cats_Manor_in_Saudi_Arabia_for_a_Western_Audience.php.
Murray, Janet H. 2012. *Inventing the Medium : Principles of Interaction Design as a Cultural Practice*. Cambridge, Mass: MIT Press.
Murray, Janet Horowitz. 1997. *Hamlet on the Holodeck: the Future of Narrative in Cyberspace*. New York: Free Press.
Pagulayan, Randy, Kevin Keeker, Ramon Romero, Dennis Wixon, and Thomas Fuller. 2009. “User-Centered Design in Games.” In *Human–Computer Interaction Handbook*, 20071544:741–59. Fundamentals, Evolving Technologies and Emerging Applications, Second Edition. CRC Press. doi:10.1201/9781410615862.ch37.
Pinchbeck, D. 2008. “Dear Esther [Video Game].” The Chinese Room
Roth, Christian, and Hartmut Koenitz. 2016. “Evaluating the User Experience of Interactive Digital Narrative.” In, 31–36. New York, New York, USA: ACM Press. doi:10.1145/2983298.2983302.
Stroebe, Wolfgang, and Fritz Strack. 2014. “The Alleged Crisis and the Illusion of Exact Replication.” *Perspectives on Psychological Science* 9 (1): 59–71. doi:10.1177/1745691613514450.
Sullivan, Gail M, and Richard Feinn. 2012. “Using Effect Size—or Why the PValue Is Not Enough.” *Journal of Graduate Medical Education* 4 (3): 279–82. doi:10.4300/JGME-D-12-00156.1.
Telltale Games. 2012. “The Walking Dead [Video Game].” San Rafael: Telltale Games.
The Fullbright Company. 2013. “Gone Home [Video Game].” Portland, OR: The Fullbright Company.