OPOS : an observation scheme for evaluating head-up play

Citation for published version (APA):

Bakker, S., Markopoulos, P., & Kort, de, Y. A. W. (2008). OPOS : an observation scheme for evaluating head-up play. In *Proceedings of the 5th Nordic conference on Human-Computer Interaction : Building Bridges, October 20-22, 2008, Lund, Sweden* (pp. 33-42). Association for Computing Machinery, Inc.
<https://doi.org/10.1145/1463160.1463165>

DOI:

[10.1145/1463160.1463165](https://doi.org/10.1145/1463160.1463165)

Document status and date:

Published: 01/01/2008

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

OPOS: an Observation Scheme for Evaluating Head-Up Play

Saskia Bakker, Panos Markopoulos,
Department of Industrial Design
Eindhoven University of Technology
P.O.Box 513
5600 MB Eindhoven, the Netherlands
P.Markopoulos@tue.nl

Yvonne de Kort
Department of Technology Management
Eindhoven University of Technology
P.O.Box 513
5600 MB Eindhoven, the Netherlands
Y.A.W.d.Kort@tue.nl

ABSTRACT

The concept of Head-Up Games [18] advocates that pervasive games of the future should be designed to evoke play patterns akin to those of traditional outdoor games. This tenet, while appealing, is ill defined without a clear description of these behaviors. Below we introduce OPOS, an observation scheme that can be used to evaluate Head-Up Games and, more generally, outdoor pervasive games intended for children. The observation scheme has been evaluated through its application in observing play with traditional outdoor games and a purpose made Head-Up Game. The study involved 24 children of 10-11 years old; it was found that the proposed observation scheme is objective and reliable, helping evaluators compare pervasive games regarding the play behaviors they provoke.

Categories and Subject Descriptors

H.1.2 [Models and Principles]: User/Machine Systems - *Human factors*. H.5.2 [Information Interfaces and Presentation (e.g. HCI)]: User Interfaces - *Evaluation/methodology, User-centered design*. K.8.0 [Personal Computing]: General – *games*.

General Terms

Design and Human Factors.

Keywords

Pervasive games, Structured observation, Game evaluation, Children.

1. INTRODUCTION

In recent years, interest has been growing towards pervasive games, an emergent genre of computer games that are intended to be played outdoors or even on the move. According to Magerkurth

et al. [10], “Pervasive games are no longer confined to the virtual domain of the computer, but integrate the physical and social aspects of the real world”.

Pervasive gaming research has spawned several concept prototypes in recent years. Earlier attempts focused on adult players, (e.g. [3, 4 and 14], but an increasing number of games targeting children have been created (e.g. [2, 6, 12 and 20]). Perhaps because of the immaturity of pervasive technology or because of the reliance upon handheld devices, the resulting game play can in many of these cases be described as ‘heads down’; players need to focus on their handheld displays. Unfortunately, paying attention to handheld devices allows for limited physical activity and prohibits the fluency characterizing traditional outdoors game-play. This critique may not apply universally or uniformly to existing pervasive games; it does however represent a serious concern when designing games appropriate for children.

In response, Soute and Markopoulos [18] articulated the notion of ‘Head-Up Games’ (HUGs), advocating the design of pervasive computer games, where interaction with technology will not inhibit physical movement and social interaction between players and would have the spontaneity that characterizes traditional playgrounds play. They argued in favor of alternative output technologies than handheld screens, and advocated the use of simple robust technologies that do not require major hardware installations bound to a very specific area (as for example, a ‘smart’ playground would be). They argue that such games can encourage open-ended play or fantasy and imagination of the players and would not require the constant attention of their users, leaving them free to interact with each other, move around, and eventually giving rise to play patterns typical for non technical outdoor play. As they put it, this should help “making the outdoor games of the future look more like the outdoor games of the past”.

While many pervasive games support aspects of HUGs, there is still very little experience in designing such games, and even less is known about whether designers can be effective in stimulating the forms of game play and social interaction envisioned. Soute and Markopoulos [18] discuss ‘Camelot’ [19] as an example of a Head-Up Game. Camelot is a mobile outdoor game designed for children of 8 to 10 years old. The game uses simple sensor technology integrated in game objects that are easily transportable. Camelot was evaluated at an elementary school, and gave rise to physical activity and social interactions as intended by its designers. However, the evaluation of Camelot, like the evaluations of related pervasive games mentioned earlier, relies on unstructured and

informal observations; claims regarding the patterns of behavior emerging remain thus only partially substantiated. An important requirement for furthering progress in this field is the development of methodology to support structured observation of game play, as an instrument to evaluate emerging play behaviors in pervasive children's games. Such an instrument will also provide a clear and operational description of the play behaviors we wish to stimulate with Head-Up Games.

Surprisingly however, we were not able to locate such an instrument or even a description of the typical behaviors in traditional outdoor games. This paper describes the development of an observation scheme describing 'Head-Up-play', which can be used for the evaluation of pervasive games, via structured observations. It is intended that a clear and operational description in the form of this observation scheme can guide both the evaluation and the design of Head-Up Games, allowing comparisons to be drawn between different games (traditional outdoor games, Head-Up Games, pervasive games or even video- or computer games). The scheme was applied to evaluate the 'LightHouse' game, a purpose-made example of a Head-Up Game.

2. OBSERVING PLAY

Methodological research for evaluating children's interactive products is a relatively young field. Many of the methods discussed in this field have evolved from methods originally aimed at the evaluation of software products for adults; see for example [11] for an extensive presentation of such methodology. Most often, observations carried out during such evaluations are unstructured, with the evaluator simply aiming to uncover areas of improvement for the interaction design that hinder the use of the product.

Most evaluations of pervasive games reported in literature are relatively informal relying on self report and unstructured observations. A rare example of a more structured study is [15] reporting the evaluation of the Hunting of the Snark, an early pervasive game for children. Their observation focused on desirable behaviours indicating playful learning, but again a structured observation scheme was not provided. In general, one can conclude that related research in human computer interaction does not provide well defined observation techniques for the evaluation of pervasive games.

Reviewing the literature in search of suitable observation based instruments for evaluating children's games, several schemes were located for observing play (e.g., [13, 16]). Most of these schemes originate in the domain of developmental psychology. Probably based on their application for the evaluation of children's development, most of them are geared towards children between 3 and 6 years old and typically focus on play rather than gaming of any kind. Such a scheme is the Play Observation Scale (POS), developed by Rubin [16].

POS distinguishes social play behaviors, cognitive play behaviors and non-play behaviors. This classification is also used in several play theories, such as described in [9]. The behaviors listed in the POS are quite broad, for example 'solitary play', 'onlooker behavior' and 'games with rules'. A similar classification is used in [7]. Such schemes can be useful when determining differences between age-groups, boys and girls, or different environments, regarding play behavior. They are however too coarse for the observation

and evaluation of game play. Many behaviors described in these schemes (e.g. 'solitary play' or 'unoccupied behavior') are likely to be sparse during game-play, while others (e.g., 'group play' or 'games with rules') will happen most of the time. Furthermore, some behaviors (e.g. 'sensorimotor play') are not likely to occur when older children are observed. Therefore, a scheme with more specific categories pertinent to the notion of outdoor play (whether this is supported by technology or not) is needed.

Another observation scheme, described by Metin [13], focuses on children between 6 and 12 years old in the playground. In [13], this scheme is used to find a relation between playground equipment and behavior and development of children. The scheme describes behavior based on specific events or actions of a child, such as 'talking', 'orienting' or 'pretending'. Although the categories in this instrument are more specific and thus more suitable for observing games, the list does not appear comprehensive and does not provide an adequate overview of all the relevant behaviors for the current purpose.

Since no existing observation scheme was found that was suitable for evaluating Head-Up-play in traditional and pervasive games, a new observation scheme was developed during an iterative development process. The process will be described in the next section (2.1); the instrument will then be presented in Section 2.2.

2.1 Observation Scheme Development

The current section will first describe the considerations regarding the structure of the instrument. It then reports on the iterations performed for the construction of a comprehensive and usable set of behavior classes and categories.

One can distinguish two approaches to structured observations, coding and describing behaviors either as events (used for event-sampling) or states (used for time-sampling) [1]. In the case of event-sampling, or a frequency-based approach [8] the observer codes and registers specific point-events, such as 'yelling'. In time-sampling, or the duration-based approach [8], actions are coded that are continuous or at least last for a period of time and therefore have a duration, for example 'running'. This type of coding results in data that describe the percentages of the total time a player was running. Depending on the type of behavior, both approaches were considered useful for observing Head-Up-play.

A second important consideration in terms of structure is related to the level at which behavior would be observed and coded, or the unit of analysis. This could be either the group, or the individual child. When looking at a group of children playing, one may be tempted to follow that part of the group that is currently active and pay little attention to the players that are temporarily not involved in the activity. This might result in a distorted or at least incomplete description of the behavior provoked by the game. It was therefore decided to score the behavior of each individual child participating in the game. This logically implies that video registration is required for observing all children in a group, as one can not reliably code and score the behavior of all of them at once. Video data shall be employed to observe multiple players.

Traditional outdoor games were used as a point of reference in identifying behavior categories for the observation scheme. First, the first author observed children, 8 to 12 years old, playing at the school playground. During breaks, there were about 150 children present at this playground. Though many different types of games were played, only structured games were observed, including hopscotch, tag, soccer, rope-skipping and hide and seek. These unstructured observations were conducted 'live' without relying on video recordings, aiming to inventorize as many relevant behavior types as possible. Five observation sessions, that each lasted 45 minutes, were carried out by one observer sitting at the side of the playground and not interfering/interacting with the children. In the first observation session, different groups of children were

observed playing different games for 5 minutes. The observer took as many notes as possible; later these notes were clustered providing a first classification of play behaviors relating to outdoor play. Based on this and the literature discussed above we created an initial version of observation scheme.

The observation scheme was improved through four successive iterations. First applying it to observe five different play sessions, and then revising it to remove ambiguities, allows the description of play behaviors observed and supports efficient coding.

After these first iterations, three independent coders applied this scheme to code to a video tape of children playing a low-tech version of the LightHouse game (described in section 3). This

Table 1. Observation scheme.

Class	Behavior	Explanation
Physical activity	Intensive physical activity	Exhausting physical activity that one can not keep doing for a long period of time. For example: running, jumping or skipping.
	Non-intensive physical activity	Physical activity that one can keep doing for a longer period of time. For example: walking, moving arms or legs, bending and standing up, crawling, moving while staying on the same location, etc.
	No physical activity	Standing, laying or sitting still. Very small movements such as coughing yawning, putting your hands in your pocket, looking at your watch, etc. while being still should also fall in this category.
Focus	Looking at other players	The player is looking at one or more other players. This does not only include looking at the face, but also looking at other parts of the body.
	Looking at game objects	The player is looking at one or more game objects. All things that are part of the game besides players and surroundings are game objects. For example a ball, a goal, a chalked circle on the ground, a hand held object, a token, etc.
	Looking at something out of sight, possibly part of the game	Looking at objects, people or surroundings that are not part of the game.
	Looking at something else	When game objects or players are out of sight of the camera and the observed player is looking in the direction of which these player(s) or object(s) likely are.
Social interaction	Functional, with another player	All interactions (verbal en nonverbal) that are functional for playing the game and directed to one or more other players or to no-one. For example instructions such as "give me the ball!", "get the monster-coin!" and "tag him!", or expressions like "John is it!", "tag!" or counting points aloud, or physical contact such as tagging, holding hand, etc that are needed to play the game.
	Non-functional positive/neutral, with another player	All interactions (verbal and nonverbal) that are not functional for playing the game, that are positive or neutral and directed to one or more other players or to no-one. For example communication about subjects that are not related to the game, showing results to other players, cheering, screaming, expressions of enjoyment and physical contact not required for playing the game such as holding hands, high five, etc.
	Non-functional negative, with another player	All interactions (verbal and nonverbal) that are not functional for playing the game, that are negative and directed to one or more other players or to no-one. For example negative communication such as swearing and bullying, expressions of pain or negative physical contact such as kicking or hitting.
	With a non-player	All interactions (verbal en nonverbal) that are directed to someone who is not a player in the game. This can be a researcher, a teacher, a parent, a peer who is watching the game, etc.
	Unintended physical contact	Physical contact that is not intended, such as accidentally bumping into another child.
General	In sight	In sight of the camera.
	Out of sight	Out of sight of the camera.

was done to evaluate the usability of this scheme and to get some feedback for improvement. The comparison of the results and a discussion between the three observers revealed some missing categories in the observation scheme, such as communication with someone who is not playing.

As a result of this iterative process, three classes of behavior of children playing outdoor games were distinguished: physical activity, social interaction and the objects or people players are looking at. Most outdoor games involve a lot of physical activity, however some also encourage standing still (e.g. hide and seek and rope skipping when waiting for ones turn). Social interaction is also an important aspect of outdoor games. Most observed social interaction is pertaining to the game (e.g. giving each other instructions or motivating each other) or needed to play the game (e.g. tagging each other). The third important aspect of outdoor games is what the players are looking at. They could be looking at other players, indicating social interaction in an indirect way, or at game objects (e.g. a soccer ball), indicating engagement in the game. These three behavioral aspects are the three main classes of the observation scheme.

2.2 The Outdoor Play Observation Scheme (OPOS)

The final observation scheme (see Table 1) includes four different classes of behavioral categories; ‘physical activity’, ‘focus’, ‘social interaction’ and ‘general’. This latter class ‘general’ does not describe aspects of play as such, but concerns practical issues, e.g., ‘in sight of the camera’ and ‘out of sight of the camera’.

The first observation class, ‘physical activity’, distinguishes intensive, non-intensive and no physical activity. Games such as soccer often result in a lot of intensive physical activity, while other games, such as hopscotch, might show more non-intensive physical activity. This distinction is made to be able to portray such differences between outdoor games.

The second class, ‘focus’, concerns what the players are looking at, namely, each other, game objects or something else. It is important to decide what the game objects are before starting the observation. They can be a soccer ball or a skipping rope, but also chalked lines on the ground, poles that serve as goals or even a researcher acting out an interactive artifact in a Wizard-of-Oz evaluation session. When part of the game takes place out of the camera view, an observer can code a player ‘looking at something out of sight, possibly part of the game’.

The class ‘social interaction’ is about social interaction with other players. This interaction can be functional for playing the game, non-functional and positive or neutral, or, finally, non-functional and negative. Players might also interact with non-players, e.g., inviting other children to play or showing achievements to an audience (which indicates engagement in the game). Interaction with an outsider to the game, (e.g., the observer), might indicate not understanding the game or not being engaged in the game. Explicitly coding this behavior helps avoid coding it as interaction with another player. The same holds for the category ‘unintended physical contact’.

The classes ‘physical activity’, ‘focus’ and ‘general’ consist of states and thus use time-sampling. The states of these classes are mutually exclusive and exhaustive. This means that a player is

always in one of those states, but never in two at the same time. For example a player is either experiencing ‘intensive physical activity’, ‘non-intensive physical activity’ or ‘no physical activity’. In contrast, the class ‘social interaction’ consists of events and thus uses event-sampling.

In order to verify that the scheme is useful in evaluating and comparing Head-Up Games (HUG’s) and traditional outdoor games, it was used to code play behavior for children playing a HUG developed ex novo, and two other existing outdoor games. First, we describe this HUG, called ‘LightHouse’.

3. LIGHTHOUSE GAME

The goal of this pirate-game is to collect as many treasures as possible. Each participant represents a pirate and has his/her own pirate ship. Physical treasures are located on the treasure islands (see Figure 1). To collect treasures, the pirates have to run to one of the treasure islands, get a treasure and return it to their pirate ship.

The treasure islands are guarded by a (physical) lighthouse (see Figure 2) that stands on a desert island. The lighthouse shines a rotating light toward the pirate ships. While collecting treasures, the pirates should not be seen by the lighthouse. If they are seen, they have to return the treasure they have in their hands and go back to their pirate ship.

While collecting treasures, the players have to be aware of a sea-monster. At random moments, the sea-monster wakes up, which is indicated by a sound. If this happens, the pirates have to go to the desert island as fast as possible. The pirate that arrives here last is ‘captured’ by the monster and half of his treasures will be taken. One of the treasures is the ‘monster-coin’. The pirate who has this treasure is friends with the monster. This means that he or she does

Figure 1. Floor plan of the lighthouse game, circles indicated with chalk.

Figure 2. Game objects in the lighthouse game; lighthouse (left) and treasures (right).

Figure 3. Impression of the lighthouse game.

not have to go to the desert island when the monster wakes up and the monster will give him or her the treasures he took from one of the other pirates. After the monster has woken up, the friend of the monster has to return the monster-coin to the treasure islands to give other pirates the chance to become friends with the monster. When no treasures are left at the treasure islands, the game ends and the pirate with the most points wins. The game can be played between individual children or teams of two. In the team version of the game, two pirates co-operate in collecting treasures. See Figure 3 for an impression of how the game is played.

4. CASE: APPLICATION OF OPOS

The OPOS is applied to guide the structured observation of children playing the LightHouse game, tag and soccer. The objective of this exercise was to assess reliability of the OPOS, practical difficulties in its application and its ability to capture differences with regards to how children play these different games. For this latter purpose, the OPOS was also applied to a short video of two children playing a video game on an Xbox games computer. This video was selected because it clearly shows behavior different from the behavior seen in the LightHouse game, tag and soccer. This section will describe the experiment setup and coding procedure. The analysis of the results is described in section 5.

4.1 Participants

The experiment was performed at an elementary school. One class of 24 children (11 girls and 13 boys, ages 10-11 years old)

participated. The class was subdivided into four groups of five, six or seven children. Boys and girls were equally divided over the groups. To make the experiment more realistic, friends were grouped together. This is common practice in evaluating games with children (see [5]).

4.2 Procedure

Two groups played the LightHouse game (one group played the individual version and the other played the team version) and two other groups played two traditional outdoor games: tag and soccer. These games were chosen as the observations during break-time showed that these were quite popular games among the target group. Furthermore, these games have some important differences: the former is an individual game that uses no game object, while the latter a team game that uses a game object (a soccer ball). It was expected that these two games should evoke very different play patterns and especially different social interactions.

During the experiment, the researcher explained the rules of the game and did not further interfere with the game, except when players initiated interaction with the researcher.

4.3 Setting and Materials

The games were played on the school playground. This is a familiar environment for playing; less familiar environments might influence variables such as enjoyment. As there was only limited time available, two games were played simultaneously on two sides of the playground.

Each game was captured on video by one camera. This camera was located at reasonable distance from the play-area, in order to capture as much of it as possible, but still close enough to capture sound as well. Screenshots of the videos taken during the experiment can be seen in Figure 4.

4.4 Coding Procedure

To evaluate the observation scheme, one coder (coder 1) coded all the behaviors of all the children in the four observed games, applying the scheme described in Section 2.2. In addition, two independent coders (coders 2 and 3) were asked to apply the scheme on a selection of the videos. The second and third coders were trained by coding a training-video, after which the differences between coders were discussed. To make sure that the selected fragments showed a variety of behaviors, one participant of the LightHouse game (female player, approximately 3,5 minutes of video), one participant of a traditional outdoor game (male player,

Figure 4. Screenshots of the videos taken during the user experiment; LightHouse game (left), soccer (middle) and tag (right).

approximately 8 minutes) and one participant of the video game (male player, approximately 3 minutes) were randomly selected. These videos were coded by both the second and third coder.

5. EVALUATION OF OPOS

The evaluation of the coding scheme consists of two parts. First it will explore its usability and inter-coder reliability using only a selection of the data gathered during the test. Second it will assess its sensitivity and utility by comparing the data for the different games observed.

5.1 Inter-Coder Reliability

In order to evaluate the different classes of the observation scheme, the inter-coder reliability has been calculated for each class, via the Kappa statistic K [17]. For these calculations, the three videos mentioned above have been interpreted as one single data set as the goal is to compare the results of the different coders rather than find differences between the games.

Jansen et al. [8] distinguish 2 methods for comparing the results of different coders.

- The first method ('tallying instances of behavior') is based on comparing the total frequency of events or the total duration of states. This method is used when the researcher is interested in summative statistics (e.g. 'what percent of the time is a player looking at other players?').
- In the second method ('labeling instances of behavior'), each individual event or state coded is compared across coders. This method is used when the researcher is interested in the sequence or timing of behaviors (e.g. 'does a player look at other players more often while running than while standing still?').

An evaluator comparing two different games might be interested in looking at overall statistics as that will give an overall impression of the behavior of the players. In that sense, the 'tallying instances of behaviors' method is sufficient for calculating the inter-coder reliability in this study. This method evaluates the extent to which different coders have coded the same frequency of events overall rather than the exact same events at the same moment. For other types of studies the exact timing of events may also be of relevance. In these cases reliability should be calculated via the method 'tallying instances of behavior'. We report both approaches in the present study. The resulting Kappa coefficients are shown in Table 2.

Results for the class 'physical activity' are shown in Figure 5. The 'tallying' method resulted in a high Kappa coefficient of 0.9. For the 'labeling' method, the total time of the coded videos

Table 2. The Kappa coefficient for each behavioral class, calculated via the methods 'tallying instances of behavior' and 'labeling instances of behavior'.

class	K (<i>'tallying instances of behavior'</i>)	K (<i>'labeling instances of behavior'</i>)
physical activity	0.90	0.70
focus	0.74	0.45
social interaction	0.73	0.24

Figure 5. The results of the three observers in the class 'physical activity'. Behaviors in average percentage of the total time.

Figure 6. The results of the three observers in the class 'focus'. Behaviors in average percentage of the total time.

Figure 7. The results of the three observers in the class 'social interaction'. Behaviors in average number of events per minute.

was subdivided into fragments of one second, which were each compared among the three coders. This method resulted in a Kappa coefficient of 0.7. The agreement between the three coders is thus very acceptable. Figure 6 shows the results for the class 'focus'. The 'tallying' method resulted in a good Kappa coefficient of 0.74. The 'labeling' method resulted in a modest Kappa coefficient of 0.45.

Figure 7 shows the results for the class ‘social interaction’. The ‘tallying’ method resulted in a Kappa coefficient of 0.73. For the ‘labeling’ method a timeframe of 3 seconds was used as a margin for deciding whether coded events by two coders were identical or not. This method resulted in Kappa coefficient of only 0.24. The agreement between the coders regarding the class social interaction is thus acceptable for the tallying method, but low for the labeling method. In contrast to the first two classes, the ‘social interaction’ class consists of events, and uses event-sampling. Since categories are not necessarily exhaustive here, the total number of coded events differs between the three coders, resulting in much lower agreement than when using the ‘tallying’ method. This explains the big difference between the two Kappa coefficients. Also, a discussion between the three coders pointed out that especially for the class social interaction, the coders experienced more problems in coding certain events. This potentially caused delays which resulted in disagreement between coding instances for the ‘labeling’ method.

From this evaluation we can conclude that, with the exception of the labeling method for social interaction events, different coders reached acceptable to high agreement after even a limited training via a training video. Agreement will likely increase with additional training and experience [1]. For example, coder 1 (who coded all the videos) could easily recognize the voices of the different participants simply because of the experience of watching all videos repeatedly. Coder 2 and 3 experienced more problems in determining which child was saying what. This probably contributed to the disagreement between coders.

5.2 Comparing Play Behaviors Evoked by the Three Games

The observed codes for the LightHouse game, the soccer game and the tag game are plotted in figures 8, 9 and 10. Where applicable, these figures show the average results of the three coders. The results for the classes ‘physical activity’ and ‘focus’ are scored and computed as average percentages of time. The results for the class ‘social interaction’ are given in average number of events per minute, as the different games have different durations. As becomes immediately clear in Figure 8, depicting physical activity behaviors, the participants in the video game are much less physically active than the participants in the four outdoor games. When only considering the outdoor games, the occurrence of ‘non-intensive physical activity’ varies most substantially between the four games. A Chi-square test shows that the four games are significantly different for this behavioral category ($\chi^2=24.3$; $df=3$; $p<.01$). The four games were also significantly different as regards ‘no physical activity’ ($\chi^2= 18.0$; $df=3$; $p<.01$). While intensive activity was roughly equal between games, children playing soccer engaged in more non-intensive activities than the others, and those in the team version of the LightHouse game engaged in less non-intensive activities.

For the class ‘focus’, Figure 9 shows that the percentage of time the participants were looking at other players was much higher in the tag game compared to the LightHouse, soccer and video games. This is obviously due to the fact that there are no game objects in a game of tag. On the contrary, the percentage of the time the participants were looking at game objects was much higher in the video game compared to the other games. When only comparing

Figure 8. The results for the five games in the class ‘physical activity’. Behaviors in average percentage of the total time.

Figure 9. The results for the five games in the class ‘focus’. Behaviors in average percentage of the total time.

Figure 10. The results for the five games in the class ‘social interaction’. Behaviors in average number of events per minute.

the soccer game and the two versions of the LightHouse game, we see that the behavior ‘looking at other players’ was significantly more frequent in the LightHouse games than in the soccer game ($\chi^2=8,3$; $df=2$; $p<0,01$). The behavior ‘looking at something out of sight, possibly part of the game’ was not seen in the LightHouse games, as the players of these games did not go out sight of the camera.

As for social interaction, the individual version of the LightHouse game evoked much less functional social interaction than the other four games did (see Figure 10), whereas the video game showed much more nonfunctional social interaction compared to the other games. Also, when comparing the remaining three games, the frequency of functional social interaction differed significantly ($\chi^2=0,12$; $df=2$; $p<0,01$). It was lowest in the soccer game, and highest in tag. Moreover, the average number of nonfunctional positive and neutral interactions was significantly different between the four games ($\chi^2=1,14$; $df=3$; $p<0,01$). This type of interaction was higher in tag than in the other games.

The analysis indicates that the observation scheme developed in the study is sufficiently sensitive to pick up differences between different types of games. The differences reported are generally in the expected directions and in line with the inherent characteristics of the games that were played. This provides a preliminary indication of the methods construct validity. The LightHouse game evoked play behaviors that differed from the behaviors in the traditional outdoor games tag and soccer.

6. DISCUSSION

The present paper describes the development and assessment of a new observation scheme for evaluating Head-Up Game behavior in traditional and pervasive outdoor games for children. The development followed an iterative procedure and was based on the observation of behaviors in traditional games, which were used here as a benchmark for evaluating Head-Up Games. The scheme described three relevant classes of behaviors in traditional play: physical activity, focus, and social interaction.

The observation instrument was evaluated in a field experiment, employing four different outdoor games played by different children from the same school, aged 10-11. Inter-coder reliability of the scheme was assessed employing two different statistical methods. Except for the labeling method of social interactions, reliability was acceptable to high. Subsequent analyses of the data of the different games indicated that the measure is sufficiently sensitive to pick up relevant differences in play behaviors between different games.

The application and evaluation of the observation instrument naturally did raise some points for discussion. First of all, the two additional coders indicated that especially behaviors in the social interaction class are hard to code as it was hard to interpret whether an event should be coded as functional or non-functional. This was manifested in the low inter-coder reliability for this class. The fact that the videos were taken from a reasonable distance, may have contributed to this ambiguity. It made it hard to hear what the participants were saying, especially if they were not facing the camera. In the future, better quality video and audio recordings would certainly add to the usability and reliability of the method. In

addition, more extensive training to become intimately acquainted with the observation categories would be beneficial.

Multiple cameras or close-up shots would also solve a second problem that was noted: that of players disappearing from sight. The LightHouse game is very location based (bound to chalked circles on the ground) and thus easy to record using only one camera. Games that are less location bound (e.g., hide and seek), or use a larger area will result in players frequently being ‘out of sight’ and ‘looking at something out of sight, possibly part of the game’. This was also seen in the soccer and tag game. For example, when the soccer ball is out of sight of the camera, the players near the ball were the most active. This activity however, could not be coded as the coder could not see these players. Occurrences like this might have influenced the results of the evaluation of the soccer and tag game. Therefore, multiple cameras should be used when evaluating games that are not very location based.

Despite this possible solution, we must conclude that some games might be less suitable to evaluate using the observation scheme than others. When thinking of games that include hiding or moving around in a large area (maybe in the woods), video recordings are difficult to make. More sophisticated recording equipment may be better suited for evaluating such games (e.g., first person recordings obtained through mobile cameras).

As mentioned earlier, players might not only be interacting with other players, but also with non-players. In the observation scheme, no distinction is made between interaction with the experimenter running the test (which might indicate that the player does not understand the game) or with an audience (which might indicate engagement). When evaluating games that are designed to evoke interaction between players and audience, it might be better to make this distinction explicit in the coding scheme. This however, could be at the cost of making the class ‘social interaction’ more complicated to code.

As explained earlier, two methods were used for calculating the inter-coder reliability; ‘tallying instances of behavior’ and ‘labeling instances of behavior’ [8]. Especially for the class ‘social interaction’, the reliability is much lower when using the latter method. This difference might be due to the implementation of the two methods or the difficulty of coding observed social interactions. The method ‘tallying instances of behavior’ results in a Kappa of 0.73, which is good, whereas ‘labeling instances of behavior’ results in a Kappa of 0.24, which is very low. This last method, using a timeframe of 3 seconds for comparing the results of different coders, seems too fine grained for this type of evaluation. However, also when looking at the results of the method ‘tallying instances of behavior’, the reliability for the class ‘social interaction’ could be improved. Once more the need for better recordings of outdoor play is underlined.

7. CONCLUSION

This paper presents an observation scheme for evaluating children’s play behavior in traditional and pervasive outdoor games. The first evaluation of this instrument indicates that it has acceptable reliability and that it seems a precise means for describing differences and similarities between different (outdoor) games regarding the play behaviors they provoke. This scheme

can especially be useful for evaluating Head-Up Games, but can also be used for comparing other pervasive games.

The main motto behind Head-Up Games is that “the outdoor games of the future should look more like the outdoor games of the past” [18]. In other words, the behavior evoked by Head-Up Games should be similar to the behavior evoked by traditional (non-technological) outdoor games. Clearly this statement is rather coarse; the analysis above showed that different traditional outdoor games provoke different behavior. The observation scheme allows making explicit comparisons between play with different games, and could help make comparisons between different Head-Up Games. Three relevant classes of behaviors have been distinguished: physical activity, focus, and social interaction. For each class, different categories of behaviors are described that typically emerge in traditional outdoor play.

As such, besides offering researchers a tool to evaluate and compare games, the observation scheme offers developers of pervasive games a design guideline. It presents an overview of the different behaviors that can be evoked by outdoor games. This overview can be useful for game designers to select intended behaviors. For evaluation purposes, researchers should choose a relevant traditional outdoor game, or class of games to compare with.

8. ACKNOWLEDGEMENTS

We would like to thank the children and teachers who participated in the experiments performed in this study. Furthermore, we would like to thank our colleagues in the Industrial Design department and Technology Management department of the Eindhoven University of Technology for their input and help in performing and evaluating the (pilot-) experiments.

9. REFERENCES

- [1] Andel, J. van. 1985. Woonomgeving en Kinderen; een onderzoek naar de invloed van ruimtelijke kenmerken en veranderingen van de woonomgeving op kinderen van 6 tot 12 jaar. Doctoral Thesis, Eindhoven University of Technology, the Netherlands.
- [2] Benford, S., Rowland, D., Flinham, M., Drozd, A., Hull, R., Reid, J., Morrison, J., and Facer, K. 2005. Life on the Edge: Supporting Collaboration in Location-Based Experiences. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Portland, Oregon, USA, April 2 – 7, 2005). CHI '05. ACM Press, New York, NY, 721-730. DOI = <http://doi.acm.org/10.1145/1054972.1055072>.
- [3] Casey, S., Kirman, B., and Rowland, D. 2007. The Gopher Game: A Social, Mobile, Locative Game with User Generated Content and Peer Review. In Proceedings of the SIGCHI International Conference on Advances in Computer Entertainment Technology (Salzburg, Austria, June 13 – 15, 2007). ACE'07. ACM Press, New York, NY, 9-16. DOI = <http://doi.acm.org/10.1145/1255047.1255050>.
- [4] Fischer, J.E., Lindt, I., and Stenros, J. 2007. Evaluation of Crossmedia Gaming Experiences in Epidemic Menace. In Proceedings of the 4th International Symposium on Pervasive Gaming Applications (Salzburg, Austria, June 11 – 12, 2007) PerGames 2007. 13 - 22.
- [5] Hanna, L., Neapolitan, D. and Ridsen, K. 2004. Evaluating computer game concepts with children. In Proceedings of the 3rd SIGCHI International Conference on Interaction Design and Children (Maryland, USA, June 1 – 3, 2004) IDC '04. ACM Press, New York, NY, 49-56. DOI = <http://doi.acm.org/10.1145/1017833.1017840>.
- [6] Harris, E., Fitzpatrick, G., Rogers, Y., Price, S., Phelps, T., and Randell, C. 2004. From Snark to Park: Lessons Learnt moving pervasive experiences from indoors to outdoors. In Proceedings of the 5th Australasian User Interface Conference (Dunedin, New Zealand, January 18 – 22, 2004) AUIC '04. 39-48.
- [7] Henniger M.L. 1980. Free-Play Behaviours of Nursery School Children in an Indoor and Outdoor Environment. In Wilkinson, P.F., Ed., In Celebration of Play: An Integrated Approach to Play and Child Development. Croom Helm Ltd, London.
- [8] Jansen, R.G., Wiertz, L.F., Meyer, E.S., and Noldus, L.P.J.J. 2003. Reliability analysis of observational data: Problems, solutions, and software implementation. Behavior Research Methods, Instruments, & Computers, 35, 3 (2003), 391-399.
- [9] Krasnor, L.R. and Pepler, D.J. 1980. The study of Children's Play: Some Suggested Future Directions. In Rubin, K.H., Ed., Children's Play: New Directions for Child Development. Jossey-Bass Inc., Publishers.
- [10] Magerkurth, C., Cheok, A.D., Mandryk, R.L. and Nilsen, T. 2005. Pervasive games: bringing computer entertainment back to the real world. ACM Computers in Entertainment, 3, 3 (July 2005), ACM press, New York. DOI = <http://doi.acm.org/10.1145/1077246.1077257>.
- [11] Markopoulos, P., Read, J.C., Hoysniemi, J., MacFarlane, S., 2008. Evaluating Children's Products: Principles and Practices for Interaction Designers. Morgan Kaufman.
- [12] Mattila, J., and Vääänen, A. 2006. UbiPlay: An Interactive Playground and Visual Programming Tools for Children. In Proceedings of the 5th SIGCHI International Conference on Interaction Design and Children (Tampere, Finland, June 7 – 9, 2006) IDC '06. ACM Press, New York, NY, 129- 136. DOI = <http://doi.acm.org/10.1145/1139073.1139106>.
- [13] Metin, P. 2003. The effects of traditional playground equipment design in children's developmental needs. Doctoral Thesis, Middle East Technical University.

- [14] Peitz, J., Saarenpää, H, and Björk, S. 2007. Insectopia – Exploring Pervasive Games through Technology already Pervasively Available. In Proceedings of the SIGCHI International Conference on Advances in Computer Entertainment Technology (Salzburg, Austria, June 13 – 15, 2007). ACE’07. ACM Press, New York, NY, 107-114. DOI = <http://doi.acm.org/10.1145/1255047.1255069>.
- [15] Price, S., Rogers, S., Scaife, M., Stanton, D., and Neale, H. 2003. Using ‘tangibles’ to promote novel forms of playful learning, *Interacting with Computers*, 15, 2, Interaction Design and Children (April 2003), 169-185.
- [16] Rubin, K.H. 2001. The Play Observation Scale (POS). Published by the Center for Children, Relationships, and Culture of the University of Maryland. Retrieved from <http://www.rubin-lab.umd.edu/Coding%20Schemes/>, last accessed 20-04-2008.
- [17] Siegel, S., and Castellan, H.J. 1988. Nonparametric statistics for the behavioral sciences. McGraw-Hill Book Company.
- [18] Soute, I. and Markopoulos, P. 2007. Head Up Games: The Games of the Future Will Look More Like the Games of the Past. In C Baranaukas et al., Ed., INTERACT 2007, LNCS 4663, Part II, 404-407.
- [19] Verhaegh, J., Soute, I., Kessels, S., and Markoploulos, P. 2006. On the design of Camelot, an outdoor game for children. In Proceedings of the 5th SIGCHI International Conference on Interaction Design and Children (Tampere, Finland, June 7 – 9, 2006) IDC ’06. ACM Press, New York, NY, 9-16.
- [20] Zuckerman, O., Pares, N., Benford, S., and Lund, H.H. 2006. Designing Interactive Environments for Outdoors Gaming and Play. SIG at the SIGCHI Conference on Human Factors in Computing Systems (Montréal, Québec, Canada, April 22 - 27, 2006). CHI ’06. ACM Press, New York, NY, 419-422. DOI = <http://doi.acm.org/10.1145/1125451.1125542>.