Data Analytics of Mobile Serious Games: Applying Bayesian Data Analysis Methods

Heide Lukosch[1], Scott Cunningham[2]
1 *Corresponding and First Author*, 2*Delft University of Technology*
[h.k.lukosch; s.cunningham]@tudelft.nl

Abstract

Traditional teaching methods in the field of resuscitation training show some limitations, while teaching the right actions in critical situations could increase the number of people saved after a cardiac arrest. For our study, we developed a mobile game to support the transfer of theoretical knowledge on resuscitation. The game has been tested at three schools of further education. A number of data has been collected from 171 players. To analyze this large data set from different sources and quality, different types of data modeling and analyses had to be applied. The application of Bayesian methods showed its usefulness in analyzing the large set of data from different sources. It revealed some interesting findings, such as that female players outperformed the male ones, and that the game fostering informal, self-directed is equally efficient as the traditional formal learning method.

Keywords: *Data analytics, Bayesian data analysis, mobile games, serious games, resuscitation*

1. Introduction

Research shows that well performed, immediate intervention of bystanders could evidently increase survival after cardiac arrest in regions with effective medicine service systems [1]. In the Netherlands, a large number of people are trained according to the guidelines of the European Resuscitation Counsel (ERC). The Dutch Heart Foundation (DHF) aims at increasing the number of citizens who are well trained to help victims of a cardiac arrest. It is argued that the increase in knowledge of treatments of cardiac arrest could decrease the mortality rate [2]. Following the promising results of studies about training school children with games, the DHF developed an app-based game together with game researchers and developers to address the training need for resuscitation. To evaluate the effectiveness of the game, we collected data of different types and from different sources, such as survey data, expert observations, and log-data. This approach allows for deep insights on the game used in the study on the one hand. On the other hand, it enables us to draw conclusions for the design of effective games aiming at the transfer of theoretical knowledge in more general terms as well as the use of new data analytics for mobile games.

1.1 Game-based learning

An increasing number of empirical studies investigates the learning effectiveness and motivational appeal of serious games. Recent meta-analyses show that serious games are effective compared to traditional instruction but that the effectiveness can still be improved [3]. Serious games can be defined in terms of being *interactive* [4], based on a set of *rules and constraints* [5], and directed towards a clear *goal* that is often set by a *challenge* [6] The 'right' level of challenge, thus the tension between a player's capability and skills, and the difficulty of game play, addresses the engagement of the player and the effectiveness of games used for learning [7]. Games constantly

provide *feedback* either as a score or as changes in the game world to enable players to monitor their progress towards the goal [4], a mechanism that can be used to foster the engagement of players.

When it comes to engagement and motivation, there often is a gap between students' daily life environment and the school environment. Nowadays, students live in a flexible and networked world with free access to the Internet and social media via smart phones, tablets and other mobile devices [8]. The school environment on the other hand is still dominated by the use of paper-based materials, theoretical lessons and fixed year groups. The curriculum usually exists of a set of fixed courses or modules with only implicit relationships between each course and module [8]. The development of the Held game, the game in our study, was conducted with the aim to overcome the traditional, less motivating way of education in school. Games offer a safe environment in which trainees are able to play, probe, make mistakes and learn [9]. Serious games make use of visual, textual and auditory channels for feedback and challenges. They enable the player to enter fantasy worlds, together with the opportunity of a strong coherence of acting as in the real world by providing rules, roles and resources [10], [11]. A game immerses players in a journey of discovery, which motivates them to learn in a playful way [12], [8]. Serious games provide access to a wide variety of online resources where traditional materials often fail [12]. They provide opportunities for social learning activities [11] and close the gap between form and content of tasks within school the setting and professional practice [8].

The need for more engaging yet at least evenly efficient resuscitation training on the one hand, and the promising results of the use of games for learning and motivating on the other, lead us to the development of the so-called Held game and the evaluation of its effects. We used the characteristics of games briefly summarized here to introduce a motivating way of resuscitation training in Dutch secondary schools, as described in section 2.

1.2 Data collection and analysis in serious games

Digital, mobile games offer the possibility to automatically collect data from the players, as they allow data to be stored, searched, retrieved and correlated [13]. Games can be used to explore and to research complex concepts and real world problems, especially when experiments in the real world would be difficult, dangerous or very expensive [14], [15]. Following [16], games can be used in research for hypothesis testing, theory construction, system specification, and phenomenological data generation. Serious games support a certain purpose beyond being only entertaining or engaging. As such, they often are included in another teaching or learning activity. Therefore, we can distinguish between data collection from the game itself, and the data collected from any activity related to the game, e.g. the debriefing, or exercises/workshops to be held in relation to the actual game play. The nature of the data collected in games can already be of different nature.

Usually, the decisions and/or actions the players make during game play are recorded, and can be correlated to the game score. Following [17], there are four types of data that van be collected from digital games: concurrent data, progressive data, longitudinal data, and external data. Concurrent data can be collected when a certain event in a game happens. In our case, when a player chooses the right action, this data is being stored. Progressive recording of data represents a string of concurrent recordings. The score at the end of one game play represents the performance of the player with regard to the actions and sequence during game play in a progressive manner. Longitudinal recording is used between game sessions and stores information for later use. For the game itself, this data is stored in order to allow for a continuation of the game play. For research purposes, longitudinal data can tell something about e.g. the retention rate of the knowledge gathered, when for example different outcomes of subsequent game sessions are compared with each other. External data is data that is collected outside of the game. In our case, we collected external data that helps us to evaluate the Held game by using a survey, expert observation, and log data from a resuscitation manikin.

The different nature of the data from the game itself, the log-data, the survey, and the expert observation, challenged us to think about an innovative way of processing the data collected. From related work, we can see that Bayesian networks are often used to predict (rational) behaviour of players in a formal game, their strategic decisions, and relation(s) to each other [18]. On the other hand, there are only very few experimental studies that examine the range of effects on learning and that those mainly use pre- and post-test-surveys only, but no formalized methods [19]. Going beyond the measurement of knowledge and skills, we are interested in a cross impact analysis of the data. Bayesian analysis solves the problems of traditional methods and

provides many advantages. There are no *p* values in Bayesian analysis, inferences provide rich and complete information regarding all the parameters, and models can be readily customized for different types of data [20].

An example of a study that examines the beliefs of players in games with incomplete information aims at mathematically formalizing the player's beliefs about each other in the game [21]. As players do make decisions in games, we can look into decision theory to find examples of how to combine different types of data related to one problem, and to analyse their relationships [22]. Thus, instead of using Bayesian networks to identify relations between players in a game, we make use of them to identify relations between data of one player, or one player group (e.g. female against male players). [23] defined the Bayesian equilibrium to be any set of mixed strategies for each type of each player, such that each type of each player would be maximizing his own expected utility given that he/she knows his/her own type but does not know the other players' types. Our own study does not measure the player's relationships in a formal game, but proposes a method to analyse the relation between the different types of data that we were able to generate within and around a mobile serious game.

The application of Bayesian models allows us to understand the relative weight of micro-level variables and macro-level or 'contextual' variables [24]. This way, we are able to analyse causal structures of the game data on one level (e.g. individual players) across a higher level of analysis (e.g. one gender). Bayesian analysis estimates directly and dynamically the value of quantity rather than the value of a test statistic [25]. In a study aiming at estimating student's evolving systems thinking skills with a game in a virtual environment, [26] used the Bayesian analysis to identify the different skills of players as well as their effects on each other. That is a way that we follow in our own study, too, in which we use Bayesian methods for data analysis. We applied three steps of data analysis to gather insights in the data, and to draw conclusions on the effect of the Held game, which is introduced in the following section.

2. The Held Game

The Held[1] Game was developed at a Dutch University together with the Dutch Heart Foundation. Several experts were involved in the process, such as an instructional expert from the health domain, and experts in resuscitation, with close contact to the national resuscitation counsel. That ensured that the game design process followed up-to-date information on the current rules and regulations in resuscitation training. Recently, resuscitation training in the Netherlands follows the standards of a national agency, and comprises out of 2 hours of theoretical knowledge transfer using a standardized power-point presentation. This presentation includes knowledge about health related issues as well as about the actions and sequences of resuscitation actions. Another 2 hours of practical training and a practical test on a mannequin follow this theoretical part. The Held game is meant to replace the first, theory related part of the training, and is followed by the same practical course unit.

To meet the target group's expectations and needs, the target group for the resuscitation course was involved in the design process from an early stage on. With an online questionnaire, a group of secondary school students (N=178) could vote for a realistic or cartoon-style design of the game. The group was also asked about scenarios and situations they would expect in a game on resuscitation. The answers revealed that the students would prefer realistic scenario's, represented in a cartoon-style design.

Starting from these design considerations, the Held game was designed along the principles of experiential learning [27]. With this design decision, we aimed at fostering the self-directed learning of students outside of formal learning contexts, and to motivate them to explore

[1] "Held" is the Dutch term for hero – players become 'resuscitation heroes' in the game.

the knowledge they need. Following up Kiili's framework of experiential learning in games, we tried to find a good balance between learning goals and gameplay, leaving enough room for challenges that foster the flow feeling of the player, resulting in a high level of engagement in the game. To allow for intrinsic motivation on the player's side, the learning environment of our game should allow students to discover new rules and ideas rather than memorizing the material that others have presented [27]. The game thus has a very strong focus on explorative actions, where students have to find out the right actions to the challenges in the game with as few instructions as possible. As we strongly believe that the physical actions to be carried out during a resuscitation situation have to be trained physically, we decided to transfer the theoretical and procedural knowledge needed into the game. For example, the frequency and depths of chest compressions as well as the right sequence of actions in a resuscitation situation is trained using the Held game. Skills that are taught during the resuscitation course, and that are represented in the game are: caring for the safety of yourself and of the victim, calling for help from bystanders and from official aid services, checking the the victim's breathing, starting CPR with breast compressions and the delivery of rescue breaths, making the victim ready for the use of an AED, and eventually using the AED. The actual practice of these actions still has to be trained in a face-to-face class, using resuscitation mannequins, as in the traditional training.

The game itself starts with a tutorial, which is obligatory when the game is played for the first time. The tutorial explains some fundamental mechanics of the game, like the hint function, and the feedback system. After that, the learning process is self-directed. The game gives immediate feedback on the actions a player makes by rewarding points, and audio-visual feedback, as highlighted in figure 1. The player can observe whether an action has been carried out correctly and whether the action took place in the right sequence.

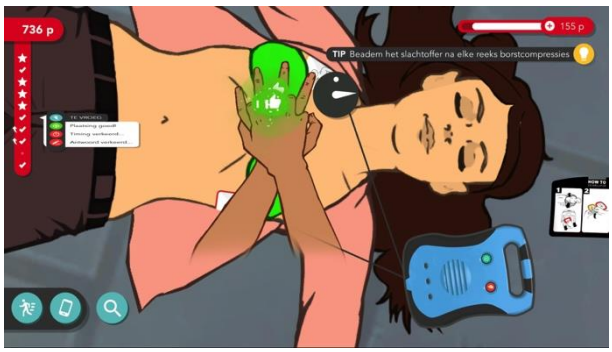

Figure 1. Interface of the Held Game, with feedback bar on the upper left corner

The Held game provides random scenarios to the player, including male and female characters with different skin colours. Scenarios take place in circumstances that are familiar to the target group, such as streets, sport fields, and situations inside of houses. All scenarios are displayed from a bird's perspective, with the ability to zoom in to the victim. In all scenarios, actions such as calling an ambulance, asking for help from bystanders, providing chest compressions, and applying an AED have to be carried out by the player in the right order. All scenarios close with arrival of the ambulance. Mini games are introduced for variation every time a player calls for the ambulance or sends a bystander to get an AED. The scoring mechanism responds to the right actions and the right sequence of actions. A player gets bonus points whenever he/she carries out the actions correctly and fast. As the game is based on self-exploratory actions, hints are only provided when a player makes the same mistakes several times, or when no action is recorded over a longer period of time. At the end of each round, the player receives an overview of his/her scores, together with the possibility to randomly unlock another hint, as shown in figure 2.

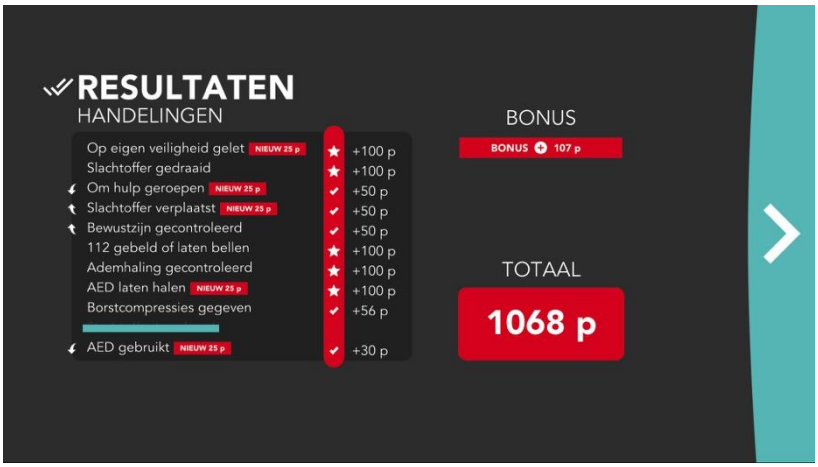

Figure 2. Overview of results in the Held game including sequence of actions, bonus points, and overall score

Players have access to a leader board including the scores of their own learning group, or a global player list. For players who are not open for competition as engaging element, we also included collectables in the form of individualized media pieces, such as fake facebook posts, newspaper articles with the players' name on it, and different types of AEDs. The idea behind these elements is to motivate the player to play as long as all elements to be collected are unlocked.

As it is the main idea to use the Held game within a formal learning setting, a portal for teachers is also developed. Through this portal, learners as well as groups of learners can be assigned to the game. Moreover, teachers can follow the learning progress of students by monitoring the number of game plays and total scores of each learner.

3. The Study – Data, Methods and Process

3.1 Material

To evaluate the effect of the Held game, we used different materials to collect data with regard to the Held game. First, we used the game itself to record the total scores of the games. Secondly, we used an online survey as knowledge test after the game play. Thirdly, we gathered information from the face-to-face training in two ways. We used a resuscitation manikin that logs data on the chest compressions and the mouth-to-mouth ventilation. Fourthly, an online survey was used to collect experiences during the resuscitation training in both groups after the full course. Finally, an online form representing a knowledge test was sent out 8 weeks after the training for testing the retention rates.

3.2 Test group

The game has been played and tested at two Dutch high schools (N = 136). At each school, several classes on different levels of education have been selected for participation in the study, based on their voluntary availability. Students were assigned to either the study group or control group by block randomization. 136 students completed the online form after the resuscitation training, and were considered for the final evaluation. The Dutch Heart Foundation initiated the participation in the study. The participating schools chose the classes and assigned the students to either the control group or the game group. Thus, the researchers were not able to influence the group division based on gender and age. The majority of the players was between 14 and 16 years old. 81 respondents were female, while 55 were female.

3.3 Process

The control group (N=69) followed a traditional ERC course, consisting out of 2 hours theoretical lessons, and 2 hours practical training. The study group (N=67) had access to the online game, and was able to play the game either on a computer or a mobile device at any time and any place. They were requested to accomplish the whole game at least 20 times within one week. Number of game play and score of every player could be monitored through the online teacher portal. The week after playing the game, the game group followed a revised ERC course of 2 hours. A pre-course questionnaire was given to half of the gaming group (N= 37). Students of both groups received a post-test survey immediately after the course. The practical test was a simulated cardiac arrest, using a Laerdal manikin. Data of the manikin, together with observations of the assessors, were collected during the test using a revised Cardiff list. Immediately after the test, all students were asked to fill-in a validated questionnaire to collect data about their experience and subjective assessment of the course. After 8 weeks, an online questionnaire has been sent out for retention test. 42 students provided answers to this questionnaire.

3.4 Data

A subset of the data was selected for further analysis. There were 136 cases, drawn upon individual student participants. Three variables were analyzed more completely in the analysis to follow.
The variables selected for further analysis include the expert rating of participant ability, the score achieved at the game, and the total number of times the game is played. Furthermore two nominal variables were created. The first includes the subset of players, which received training with the simulation game. In the sample 67 received training with the simulation game, and 69 did not receive training with the simulation game. The second nominal variable was the gender of the participant. In this sample 81 of the participants were female, and 55 were male.

The variables used in the analysis are shown below in table 1, descriptive statistics. The table lists the variables, how and if the variables are further transformed, the scale of the variables, and whether the variable is used as a dependent or independent variable in the analyses, which follow.

The game score displayed the progressive data of the game play conducted by the players. Only the numeric overall score was taken into account for our study. The survey after the resuscitation training included 27 questions, collecting demographical data, data with regard to the content of the training (5-point Likert scale), data based on a game experience questionnaire [28], and three open questions with regard to the game content. The expert observation followed a validated evaluation method in resuscitation, the Cardiff list. The list contains 17 items to measure the skills shown during resuscitation training. During the practice phase of the resuscitation training, data from the manikin was collected. The manikin automatically measures how often and how deep the chest is compressed as well as how often and how deep ventilation is done. The knowledge test after 12 weeks contained 14 questions, with 4 collecting statistical data on the participant as well as on the mode of training (traditional or game-based). 10 questions tested the knowledge of the players with the use of multiple-choice questions with 4 possibilities each, including one right answer.

Table 1. Descriptive Statistics

		Link	Predictors	Scale	N.A.	Min.	25%	50%	75%	Max.	S.D.
All (N= 136)	Expert rating	Linear	In-dependent	Ordinal	0	6.0	12.0	14.0	15.0	17.0	2.48
	Game score	Linear		Ratio	69	3782	6463	9257	15760	101030	2088.5
	Game score	Logarithm	In-dependent	Ratio	69	8.2	8.8	9.3	9.7	11.5	0.7814
	Previous experience	Linear	Dependent	Ordinal	25	2.0	4.0	5.0	6.0	8.0	1.45
	Times played	Linear		Ratio	69	9.0	20.0	22.0	32.0	126.0	2.41
	Times played	Logarithm	Dependent	Ratio	69	2.20	3.00	3.09	3.47	4.84	0.5256
Females (N= 81)	Expert rating	Linear	In-dependent	Ordinal	0	10.0	13.0	14.0	15.0	17.0	2.48
	Game score	Linear		Ratio	45	3782	7675	9623	15383	101030	2088.5
	Game score	Logarithm	In-dependent	Ratio	45	8.24	8.95	9.17	9.64	11.52	0.7814
	Previous experience	Linear	Dependent	Ordinal	13	2.0	4.0	5.0	6.0	8.0	1.45
	Times played	Linear		Ratio	45	9.0	20.0	22.5	29.00	122.0	2.41
	Times played	Logarithm	Dependent	Ratio	45	2.20	3.00	3.11	3.37	4.80	0.5256
Males (N= 55)	Expert rating	Linear	In-dependent	Ordinal	0	6.0	11.0	14.0	15.0	17.0	2.65
	Game score	Linear		Ratio	24	3973	5414	7826	16513	100466	2088.5
	Game score	Logarithm	In-dependent	Ratio	24	8.28	8.60	8.97	9.71	11.52	0.7814
	Previous experience	Linear	Dependent	Ordinal	12	2.0	4.5	5.0	6.0	9.0	1.45
	Times played	Linear		Ratio	24	13.0	20.0	21.0	36.0	126.0	2.41
	Times played	Logarithm	Dependent	Ratio	24	2.57	3.00	3.05	3.58	4.84	0.5256

3.5 Methods

Descriptive statistics, as shown in table 1, is useful to get an overview of the data. However additional analyses are needed to examine the effects of gender, training, game play, and prior experience with computers on the outcome of the game. Most importantly analyses may be used to gain additional insight on whether the game actually improved resuscitation and first aid skills.

For these purposes we choose a Bayesian analysis. Bayesian analysis presents a departure from classical statistical methods in four distinct ways. Many Bayesian analysts interpret probabilities in a different manner than classical (or "frequentist") statistics. Secondly many Bayesian analysts have different assumptions concerning the nature of models and parameters. Thirdly many Bayesian analysis interpret their statistical results differently than a frequentist analyst.

For many Bayesians probabilities are subjective statements of belief, which are conditioned on the intents, purposes and conduct of the analysis itself. This is different than a frequentist statistician who interprets probabilities as "long term" or true expressions of a real-world phenomenon. In this analysis we employ a Bayesian statistical procedure without introducing strong prior hypotheses concerning our expected results. The resultant models are known as empirical Bayesian models as they are entirely data (and not belief) driven.

In Bayesian models all parameters are stochastic. This assumption recognizes that in many settings there are multiple, uncontrolled influences which potentially affect our results. In this simulation gaming context for instance, the motivation of the student is an external factor which cannot be readily measured or controlled. Nonetheless this external variable has a strong effect on the effectiveness of training and the effectiveness of the simulation game.

The use and inference of models also differs when adopting a Bayesian methodology. The posterior density of parameters reflects expected variation in parameters across future experiments. So for instance if the density of the posterior distribution is substantially greater than zero, then we may conclude that with future experiments the parameter is equally likely to remain above zero. This intuitive characterization of the prediction interval is an appealing feature of the Bayesian approach. Despite the frequent misinterpretation of frequentist confidence intervals in this manner, this is not the appropriate interpretation of a frequentist statistical test. Bayesian analyses report the highest density interval, which is the least area under the posterior distribution which contains 95% of the mass of the distribution.

The Bayesian approach clearly distinguishes decision-making significance of the result. This requires a statement of "regions of practical equivalence" or ROPE values. Classical statisticians consider the power of their statistical tests to determine whether the results are significant for decision-making. A variety of different criteria are used in this study when setting the ROPE. These are listed below, and the resultant ROPEs are discussed in more detail in the analyses which follow.

Table 2. Criteria used in Setting the ROPE

	Criteria Considered in Setting the ROPE
Expert Evaluation	• Scores needed to pass the examination • Standard deviations of results
Game Performance	• Differences between quartiles in performances • Standard deviation of results • Absolute percentages in game performance
Learning Effects	• Amount of play required to attain significant differences in expert assessment

4. Data Analysis

As shown above, we used different materials providing several types of data in our study with the mobile game Held. For being able to evaluate the effects of the game, all data has to be considered and processed. As of the diverse nature of the data collected, we cannot rely on traditional and simple ways of data processing, but have to develop new ways of data analysis. In the following

sections, we describe a three-step approach towards the data analysis, including the results of each step and a brief discussion.

All models below are implemented using JAGS software, using an R interface. This software enables the easy creation and computer experimentation of a wide range of possible statistical models. This open source software is consequently of significant value in estimating Bayesian models (c.f. [29]).

4.1 Analysis 1

The first analysis examines performance at heart resuscitation, as evaluated by experts. The purpose of the analysis is to evaluate whether there are systematic differences in expert assessment by groups. Understanding these systematic differences, if they exist, is necessary to better evaluate the role of the game in training.

As noted the expert opinion protocol rates participants on 17 different tasks involved in heart resuscitation. Scores can vary from 17 (all tasks performed correctly), to 0 (no tasks performed correctly). Normality assumptions are not appropriate here because the model will estimate responses above and below the known scoring thresholds.

Model

Because of these data scaling issues, a novel probability model based on the Beta distribution, is used. The Beta distribution has two shape parameters. The first parameter, here called alpha, shifts the distribution rightward towards more correct answers. The second parameter, here called beta, shifts the distribution leftward towards fewer correct answers. As alpha and beta increase in value the model predicts more confidence in scoring outcomes. On the other hand, values near zero predict more widely distributed results.

The analytical model is shown below. The predicted value (number of correct performances by expert rating) is given by y1. The shape parameters, as described above, are given by alpha and beta. In turn these parameters are predicted by a baseline effect (subscripted with 0), and two nominal effects.

$$y_1 \sim 17 * \text{Beta}(\alpha, \beta)$$
$$\alpha = \alpha_0 + \vec{\alpha}_1 + \vec{\alpha}_2$$
$$\beta = \beta_0 + \vec{\beta}_1 + \vec{\beta}_2$$

Two group effects are hypothesized. A test to see whether there are gender effects is incorporated in the model, and a test to see whether there are effects of the pre-training with the game. These nominal variables are noted with an arrow. The gender effects are subscripted with 1, and the pre-training effects are subscripted with a 2.

The resultant model is related to the classical ANOVA model. There are two major differences. The first difference is that the model uses a beta distribution rather than a normal distribution, for reasons described above. Secondly the baseline performance given by experts is estimated by the model. This is different than an ANOVA model, where this parameter is set deterministically according to the data using population means.

The resultant Bayesian ANOVA model requires prior probabilities to be set. All model parameters are set using uninformative priors. The data itself then becomes the primary determinant of model outcomes, rather than initial beliefs. All priors are set to a uniform distribution, ranging from zero to one.

$$\alpha_0, \vec{\alpha}_1, \vec{\alpha}_2 \sim \text{Uniform}(0, 10)$$
$$\beta_0, \vec{\beta}_1, \vec{\beta}_2 \sim \text{Uniform}(0, 10)$$

The resultant model (implemented in R, using jags) convergences rapidly with no evidence of problems with the Monte Carlo Markov Chain (MCMC) procedures used in the estimation.

The R2 of the model is 34.4%. This is a low number, but the model itself does not include individual differences, previous experience and training, and educational program. These factors

are undoubtedly a large portion of the unexplained variance. Nonetheless the purpose of the model is not to predict expert assessments, but to better understand sources of systematic differences across experts. This requires more complete diagnostics, and requires examination of additional characteristics of the data, such as the posterior predictive distribution and contrasts between participating groups.

Results

There are six different parameters, each with associated posterior distributions that are associated with this model. Rather than to explore each of these distributions in detail, the two main contrasts are reported instead. This is followed with inference procedures to better understand the magnitude and significance of the reported effects.

Contrast one, between male and female respondents, is estimated using the baseline parameter and the associated alpha and beta parameters. The contrast is expressed as the difference in rating between women and men. For instance, a positive number indicates the number of performance assessments a man is expected to score more highly than a woman. Contrast two, between game playing and regular participants, is also estimated using the baseline parameters and the associated alpha and beta parameters. The contrast is expressed as the difference between non-game playing and game playing participants. Thus, positive number indicates that the game-playing participant performed better under expert assessment than the non-game playing participant.

Estimated contrasts are given in figure 1 below. Bayesian models are often estimated computationally rather than analytically, so the results have to be generated by computer experimentation. Nonetheless the results can be estimated to arbitrary precision given sufficient computational time. The weight under the posterior predictive distributions below is the number of times the result was simulated out of a total of 100,000 model runs.

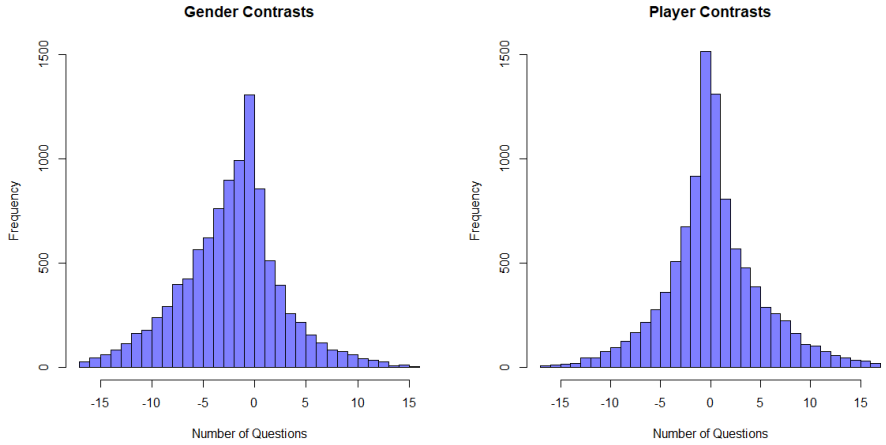

Figure 3. Estimated Contrasts from Model 1

For contrast 1, measuring systematic differences in assessment by gender, the posterior predictive distribution skews left. This indicates a relatively large effect where by experts rate female participants lower – and often by a substantial amount. For contrast 2, there are often no significant differences between game playing and non-game playing assessments by experts. The distribution may skew slightly in favor of non-game playing participants. Further diagnostic measures can help clarify these systematic differences in assessment.

Two diagnostic measures often used on posterior predictive distributions are the highest density interval (HDI) and the region of practical equivalence (ROPE). The HDI provides an indication of credible parameter ranges. It is measured by the largest area under the curve containing 95% of the probability mass. The HDI is the equivalent of confidence intervals calculated using standard statistical procedures.

Table 3. HDI and ROPE Values for the Model

	Contrast 1	Contrast 2
HDI	[-11.0, -2.2, 5.8]	[-7.8, 0.1, 8.5]
ROPE	[46.6%, 37.6%, 14.9%]	[25.8%, 47.7%, 26.3%]

These results suggest that women on average, are assessed between 11 items lower and up to six items higher than men. The average is 2.2 items lower on average. The non-game playing subgroup scored up to 7.8 items lower than the game playing group, and on the high end up to 8.5 questions higher. The extra weight in the right hand part of the tail indicates that when non-game players do better, they do better across multiple items. Nonetheless the actual difference between groups is very low; 0.1 items different in favor of the non-game playing group.

The ROPE helps establish the policy or decision-making significance of the results. A swing of performance of up to 25% of the assessment would be a significant magnitude. This is equivalent to passing or failing 4.25 of the total performance assessments. Nearly 47% of women fall below these ranges on the negative side, while only 15% score higher than these ranges on the positive side. In contrast, with the non-game playing group 26% score below the mean, and another 26% score above the mean.

Discussion

The Bayesian ANOVA model demonstrated in this section reveals substantial, significant, but unexplained negative performance by female participants in heart resuscitation training as evaluated by experts. The other contrast made is between game playing and non-game playing training participants. There is very little evidence that the pre-game training affected expert assessment of performance either for the better or for the worse.

4.2 Analysis 2

The second analysis examines performance as assessed by experts, but now examines whether or not performance at the game results in higher expert assessment. The presence of a learning effect, whereby high scoring participants also were highly assessed by experts, may provide evidence of the beneficial effects of simulation games in this training setting. For this analysis only the cases where participants played the simulation game are used. In all cases the participants were given expert assessment of resuscitation skills.

Model

A standard log-linear model relates the predicted variable (expert assessment score) and a single predicted variable (the score at the game). This is a general linear model, and like other general linear models it involves link functions, and a probability distribution as output. A logarithm is used as a linkage function. The output is a normally distributed variable.

$$y_1 \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu = b_0 + b_1 \cdot \log(x_1)$$

This is a Bayesian model, so that the parameters are stochastic. Priors are needed for a Bayesian model. The priors for the model are shown below. The model is robust to these choices, as expected for an uninformed prior.

$$b_0 = \mathcal{N}(0, 10)$$
$$b_1 = \mathcal{N}(0, 10)$$
$$\sigma = \text{Uniform}(0.001, 100)$$

Figure 2 shows the posterior distribution for three parameters of the model. These are b0, b1 and sigma. The corresponding plots are labeled the slope, the intercept, and R-squared. The

sigma parameter is recoded as an R-squared value. In conventional regression models the R-squared is a constant value. In the Bayesian models the R-squared is stochastic, and potentially assumes a range of values.

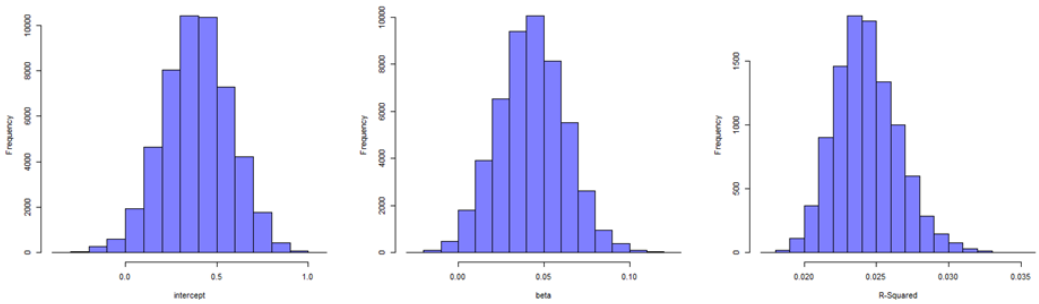

Figure 4. Estimated Parameters from Model 2

Results

In the next paragraphs the posterior predicted distribution is interpreted. A positive beta parameter is associated with game-related learning. In other words, the higher the simulation game score the higher the expert rating. The beta parameter is greater than zero 99.1% of the time, indicating highly credible evidence of game-related learning. Unfortunately though the magnitude of the effect is quite small. The average participant scores one question greater out of 17 assessment items.

Consequently the R-squared, or explained variance, on the model is also quite low. It shows a peak of 2.5% R-squared, a very low proportion of variance. As noted in the first analysis most of the variance in expert assessment remains unexplained. It may be related to individual performance, educational qualification of participants, or differences in ratings between expert assessors.

Discussion

The second analysis demonstrates consistent learning effects for participants in the simulation game. A play of the game results in one extra item rated positively by experts, out of a total of seventeen assessed items of resuscitation performance. This is a small, yet unambiguously positive learning effect for participants in the mobile game.

> The analysis demonstrated learning among the gaming participants, all other effects considered. The objective of the analysis is not necessarily to explain why certain individuals achieved the expert ratings that they did. Thus the relatively low R-squared of the model is not a detriment. The third and final analysis focuses entirely on inter-group differences in game performance.

4.3 Analysis 3

The following analysis focuses exclusively on the computer game score of participants. This analysis demonstrates significant differences in learning by men and women who play the game. All things considered women perform better at the game than men, but men show a greater upside in performance. Men learn slightly more over repeated plays, and are better able to apply previous knowledge with computer games when playing the game. While the baseline differences in performance are significant, the differences in learning between the two groups are probably not significant.

Model

The model formulates the final score of the game as a function of one nominal variable (the gender of the participant), and two metric variables (the number of times played, and self-professed experience with computer games). The resultant model is an ANCOVA model. However it differs from a classic ANCOVA model given the Bayesian formulation as shown below.

$$\log(y_2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$
$$\mu_2 = b_{20} + b_{21} \cdot \vec{x}_1 + m_1 \cdot \log(x_2) + m_2 \cdot x_3$$

The variables in the model are the game score (y2), the gender of the participant (x1), the number of times played (x2), and previous experience with gaming (x3). Non-informative priors were set for all parameters. The model is robust to these assumptions.

Results

The baseline performance of the game is shown below by gender. Note that this baseline conditions on previous computer experience by the participant, and the number of times the game is played. Male baseline performance (at left) is lower than female performance (center). Difference between the groups is shown most clearly at right, in a contrast which shows the added performance of female participants over male performance.

The highest density interval (HDI) on the posterior distribution ranges from -0.38 to 1.00. Given this distribution female participants on average scored 0.39 units higher than men. Since this is on a logarithmic scale, this indicates an average of 48% higher scores. We may further conclude that the game is particularly beneficial for the female participants. The effect is robust as nearly 80% of the mass density is higher than zero.

Another issue is whether these differences are meaningful in training terms. The region of project equivalence (ROPE) ranges from -0.666 to 0.666. Almost 99% of the mass of the distribution is higher than -0.666. Some 20% is higher still than 0.666. Both of these factors suggest a very significant performance bonus attained by female players of the game.

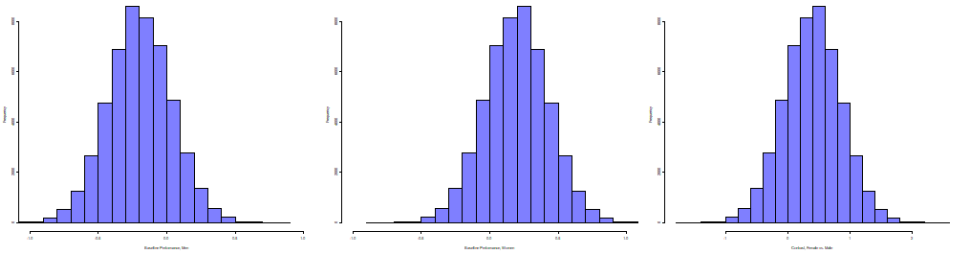

Figure 5. Baseline Performance by Gender

Learning effects are now discussed, and shown in the figure below. The region of practical equivalence (ROPE) is one or below. Learning effects less then this would suggest that most players do not gain mastery of the game in a reasonable about of plays. In this case all the mass of the posterior distribution is greater than 1.0, suggesting significant amounts of learning. Somewhat more learning is gained per play of the game by men, then by women (0.111 units). The difference is not practically significant, and these differences may be solely the result of chance. Previous experience at other computer games is somewhat detrimental for both men and women. Here again the results are not practically significant, and are results which could have occurred by chance.

ISSN: 2384-8766 http://dx.doi.org/10.17083/ijsg.v5i1.222

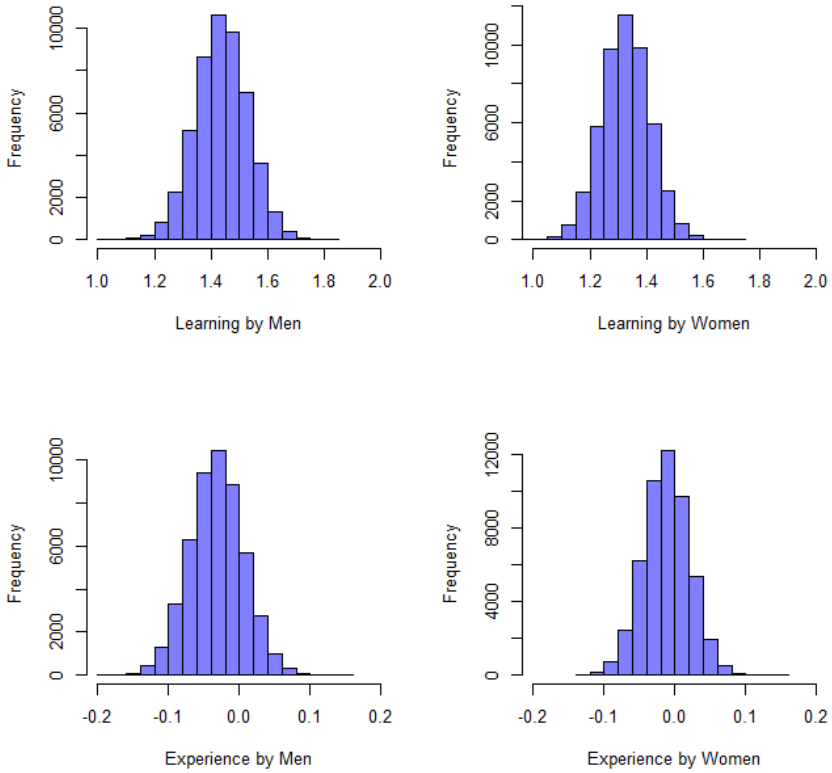

Figure 6. Learning and Experience, by Gender

As a final element of this analysis we examine the predicted variance, or the equivalent R-squared measure. The model explains some 55 to 60% of the total variance in the data. The number of plays of the game, in particular, is strongly correlated with the final score earned in the game.

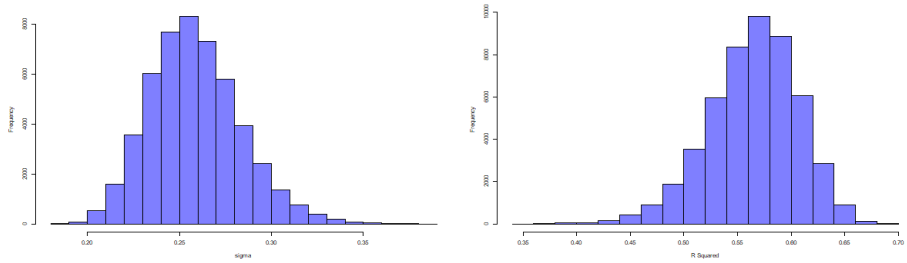

Figure 7. Explained Variance in the Model

Alternative variants of the analysis were explored involving heterogeneous sources of variance, with and without experience, and with and without metric variables. This formulation is chosen because of the substance and predictive character of repeated play, and because of the interesting interaction between baseline and learned performance. These results demonstrate that female participants demonstrated an unusual aptitude for the game. The results are robust to chance, and demonstrate little or no possibility of a negligible effect on female players. The result is significant in part because previous studies have suggested female participants are often at a disadvantage when playing computer games or simulations. Previous lack of experience with computer games does not substantially hurt performance at the game. Likewise there is a strong

effect on learning given repeated plays, and little or no evidence of differences in learning across repeated plays between female and male participants.

5. Conclusion and Discussion

The female players are at no disadvantage when playing the computer simulation. In fact there is some evidence that female players are superior to male players, achieving a higher score prior to repeated exposure and play. Furthermore the game may be more motivational for female players given the systematically worse assessments given to female resuscitation students by their instructors.

The expert assessments of resuscitation may be biased. All things considered high performance at the game leads to a slightly higher expert rating. Furthermore the costs of the game in terms of alternative pedagogical options are pretty mild. A segment of students fail to get the highest possible assessments from experts. This might be ok since the expert rating is not necessarily the gold standard for actual performance. Controlled for computer experience, the female players outperformed the male ones. It may be the case that computer experience hindered the male players to perform better. In another study on the performance of players in a digital Microgame focusing on integrated planning skills [30], we were also able to find a better performance of the female players. Usually, literature reports that male participants perform better in virtual or digital environments, explained by a higher use of such environments in the male population, and a higher anxiety level in the female user group [31], [32], [33]. Explanations also state that often, content of games addresses male players in favor to female players [34]. As we considered the requirements of as well male as female players from the beginning of our design process onwards, and aimed at a gender and culture sensitive design of the interface, we might have been able to limit these influences in our game.

In our article, we showed why and how we applied new ways of data analytics to evaluate the effects of a mobile game used in resuscitation training. We illustrated that we had to look for innovative ways of data analysis as our evaluation consisted of various types of data, from numerical to descriptive ones. By applying Bayesian data analysis, were able to find that

- Expert observers systematically and severely underrate the skills of female participants in resuscitation training,
- There is a substantial relationship between game score and expert rating,
- Female participants with computer gaming experience out-perform males in the game,
- The game session limits the ability of some participants to gain the highest expert scores.

The analyses of the data show a gender bias in the skill evaluation of expert observers in favor of the male participants. Further studies report on a poor correlation between the use of the checklist we applied in our study and the recording of the manikin [35], [36]. The gender bias problem obviously is not new, as it is also reported in other than game related domains (see e.g. [37]). This indicates the necessity of applying a combination of evaluation methods on the effects of games to avoid subjective indication of performances. Expert assessments may be biased since all things considered high performance at the game leads to a higher expert rating. With some critical distance given the gender bias and the studies on the mismatch between observer ratings and more objective data found, we can conclude that a good performance in the game lead to a good performance in the practical training part, as recorded by the expert observations. That indicates a positive effect of the mobile game Held on resuscitation skills.

5.1 Results of the game

The 10 knowledge-related questions of the pre-course questionnaire showed a low rate of incorrect answers (m = 5.6), with the highest percentage of wrong responses with regard to the use of the AED (n = 10), the connection to the centralist (n = 9) and the right actions to control the breathing of the victim (n= 18). This supports the concept of the new, game-based resuscitation course as a combined course of knowledge transfer by the game and face-to-face practical training.

Together with the observations and feedback from the Laerdal manikin, we can conclude that the Held game is a capable tool of knowledge transfer within an ERC course with certain limitations. We are aware of the fact that the game cannot address the use of the AED in sufficient detail, as well as the correct action of how to check for breathing and to conduct chest

compression. The second part of the game-based ERC course should thus focus strongly on practical actions, while the game itself can be used for knowledge transfer in the first place.

With regard to the cartoon-like design of the Held game, based on the preferences of the target group tested in an early design stage, we can conclude that this style supports the positive effects of the game. A moderate analysis of visual styles in serious games shows that instructional techniques in serious games with basic and cartoon-like visual representations are more effective than instructional techniques in serious games with a (photo)-realistic design [3]. Following [3], these findings are corroborated by other reviews [38], [3]. A possible explanation is that schematic/cartoon-like designs facilitate students to focus on relevant information in the game, whereas in games with (photo)-realistic designs students can easily become overwhelmed by the visual complexity [3]. The visuals in the Held game are indeed simple, yet appealing, representing a simplified version of a real (and possible) situation where resuscitation might or might not be the right action to take. Based on these results, we recommend the use of cartoon-like design for mobile serious games.

5.2 Results regarding the method used

A Bayesian analysis method presents several advantages for the analysis of simulation games and participant learning. The technique is designed for the analysis of open and uncertain systems. This certainly applies to the serious gaming setting where there are multiple sources of incomplete, uncontrolled and uncertain factors, which have a strong effect on the results.

Bayesian analysis techniques are sometimes critiqued for their subjective character. While it can be an advantage to introduce prior hypotheses (based for instance on prior studies of learning, simulation or gaming), some would critique this as being overly subjective and therefore subject to experimental vagary. This study demonstrates the benefits of a Bayesian approach while using flat, uniformed priors. This approach is also known as the empirical Bayes method. Furthermore, the Bayesian technique can be implemented using flexible, open source tools. These probabilistic programming tools afford analysts the capability to develop the appropriate models for the setting. This allows for the design of novel, ecologically valid analysis techniques such as the Beta ANOVA example discussed earlier.

A final advantage of the Bayesian results is the support it offers for constructive inference on the data. Bayesian inverse probabilities are easy to interpret. Furthermore it is straightforward to judge how credible it is that model parameters could reach or exceed certain variables in repeated testing or replication. The magnitude of the results are easily and explicitly assessed using concepts such as the region of practical equivalence (ROPE). These quantities can be assessed using domain knowledge. Such knowledge may be readily gathered as a central part of the game design efforts. As the application of Bayesian data analysis methods enabled us to gather a deeper understanding of the effects of the game on different learner groups, and insights in the effects of different evaluation methods, we can recommend the use of this data analysis technique in game-related studies including data from different sources.

5.3 Limitations of the study and future research

One limitation of our study is that we were not able to select the user group, as the participating schools assigned the students based on availability and scheduling. Still, as we recruited classes that represent the actual target group for the game, we have been able to set up an at least quasi-experimental setting for our study. Based on our experience with data analytics, we were able to develop new ways of data processing, using Bayesian methods that revealed surprising findings with regard to the influence of the evaluation methods on the outcomes of the game. Future research will aim at applying even further analytical methods to test the data, and will include further populations to validate the findings of our study. We will apply our participatory design approach to further games, to test whether we can generalize the positive effects of gender and culture sensitive design as well as the use of cartoon styled interfaces for serious mobile games. This will also contribute to the development of a toolkit on mixed methods of data analytics for games.

Acknowledgements

This study has been realized with support of the Dutch Heart Foundation. Simon Tiemersma of the TU Delft gamelab and Olivier Hokke of studio Wolfox are the main developers of the Held game.

References

[1] Holmberg, M., Holmberg, S., & Herlitz, J. "Factors modifying the effect of bystander cardiopulmonary resuscitation on survival in out-of-hospital cardiac arrest patients in Sweden". European Heart journal, 22(6), 511-519, 2001. https://doi.org/10.1053/euhj.2000.2421
[2] Rea, T. D., Pearce, R. M., Raghunathan, T. E., Lemaitre, R. N., Sotoodehnia, N., Jouven, X., & Siscovick, D. S. "Incidence of out-of-hospital cardiac arrest". The American journal of cardiology, 93(12), 1455-1460, 2004. https://doi.org/10.1016/j.amjcard.2004.03.002
[3] Wouters, P., & van Oostendorp, H. "Overview of Instructional Techniques to Facilitate Learning and Motivation of Serious Games", In Instructional Techniques to Facilitate Learning and Motivation of Serious Games, 1-16. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-39298-1
[4] Prensky, M. Digital game-based learning. New York: McGraw-Hill, 2001.
[5] Garris, R., Ahlers, R., & Driskell, J. E. "Games, motivation, and learning: A research and practice model", Simulation and Gaming, 33, 441–467, 2002. https://doi.org/10.1177/1046878102238607
[6] Malone, T. "Toward a theory of intrinsically motivating instruction". Cognitive Science, 4, 333–369, 1981. https://doi.org/10.1207/s15516709cog0504_2
[7] Kiili, K., de Freitas, S., Arnab, S., & Lainema, T. "The design principles for flow experience in educational games". Procedia Computer Science, 15, 78-91, 2012. https://doi.org/10.1016/j.procs.2012.10.060
[8] Veen, W., & Vrakking, B. "Homo Zappiens. Growing up in a digital age", London: Network Continuum Education, 2006.
[9] Shaffer, D. W. How Computer Games Help Children learn. New York: Palgrave Macmillan, 2006. https://doi.org/10.1057/9780230601994
[10] Chalmers, A., & Debattista, K. "Level of Realism for Serious games". In IEEE Proceedings of 2009 Conference in Games and Virtual Worlds for Serious Applications, 225-232. Coventry: IEEE, 2009. https://doi.org/10.1109/VS-GAMES.2009.43
[11] Klabbers, J. H. The Magic Circle: Principles of Gaming & Simulation. Rotterdam/Taipei: Sense Publishers, 2009.
[12] Gee, J. P. What video games have to teach us about learning and literacy. Computers in Entertainment (CIE), 1(1), 20-20, 2003. https://doi.org/10.1145/950566.950595
[13] Murray, J. H. Hamlet on the holodeck: The future of narrative in cyberspace. MIT press, 2017.
[14] Raser, J. R. Simulation and society: An exploration of scientific gaming. Allyn and Bacon, Boston, MA, 1969.
[15] Duke, R. D., & Geurts, J. Policy games for strategic management. Rozenberg Publishers, Amsterdam, 2004.
[16] Greenblat, C. S. “Gaming-simulation as a tool for social research”. In Greenblat, C. S. and Duke, R. D. (Eds.) Gaming-simulation: Rationale, design, and applications. Sage Publications, New York, NY, 320–333, 1975.
[17] Medler, B. “Generations of game analytics, achievements and high scores”. Eludamos. Journal for Computer Game Culture, 3(2), 177-194, 2009.
[18] Brynielsson, Joel, and Stefan Arnborg. "Bayesian games for threat prediction and situation analysis." Seventh International Conference in Information Fusion, Stockholm, Sweden, June 28 – July, 1, 2004.
[19] Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. “Melding the power of serious games and embedded assessment to monitor and foster learning”. In: Ritterfeld, U., Cody, M., & Vorderer, P. (eds.) Serious games: Mechanisms and effects, 2, 295-321, 2009.

[20] [20] Kruschke, John K. "What to believe: Bayesian methods for data analysis." Trends in cognitive sciences 14(7), 293-300, 2010. https://doi.org/10.1016/j.tics.2010.05.001
[21] Mertens, J.-F., & Zamir, S. "Formulation of Bayesian analysis for games with incomplete information." International Journal of Game Theory, 14(1), 1-29. https://doi.org/10.1007/BF01770224
[22] Berger, J.O. "Statistical decision theory and Bayesian analysis". New York: Springer, 2013.
[23] Harsanyi, J.C. "Games with incomplete information played by 'Bayesian' players. The basic model". Management Science, 14(3), 159–182, 1967. https://doi.org/10.1287/mnsc.14.3.159
[24] Jackman, S. "Bayesian analysis for the social sciences". Vol. 846. Chichester: John Wiley & Sons, 2009. https://doi.org/10.1002/9780470686621
[25] Reckhow, K. H. "Bayesian approaches in ecological analysis and modelling". In C. D. Canham, J. J. Cole, & W. K. Lauenroth (Eds.), The Role of Models in Ecosystem Science, 168–183. Princeton, NJ: Princeton university Press, 2002.
[26] Shute, Valerie J., Masduki, I. & Donmez, O.. "Conceptual framework for modeling, assessing and supporting competencies within game environments." Technology, Instruction, Cognition & Learning 8(2) (2010).
[27] Kiili, K. "Digital game-based learning: Towards an experiential gaming model." The Internet and higher education, 8(1), 13-24, 2005. https://doi.org/10.1016/j.iheduc.2004.12.001
[28] IJsselsteijn, W., De Kort, Y. A. W., & Poels, K. "The game experience questionnaire". Manuscript in preparation, 2008.
[29] Plummer, M. "JAGS: A program for Bayesian analysis of graphical models using Gibbs sampling". CiteSeer. doi: 10.1.1.13.3406, 2003.
[30] Lukosch, H., Kurapati, S., Groen, D., & Verbraeck, A. "Gender and Cultural Differences in Game-Based Learning Experiences". The Electronic Journal of e-Learning, 15(4), 310-319, 2017.
[31] Bryant, K. J. "Personality correlates of sense of direction and geographic orientation". In Journal of Personality and Social Psychology, 43(6), p. 1318, 1982. https://doi.org/10.1037/0022-3514.43.6.1318
[32] Lawton, C. A. "Gender differences in way-finding strategies: Relationship to spatial ability and spatial anxiety". Sex Roles, 30(11-12), 765-779, 1994. https://doi.org/10.1007/BF01544230
[33] Lawton, C. A. "Strategies for indoor wayfinding: The role of orientation". Journal of Environmental Psychology, 16(2), 137-145, 1996. https://doi.org/10.1006/jevp.1996.0011
[34] Kafai, Y. B., Heeter, C., Denner, J., & Sun, J. Y. "Beyond Barbie [R] and Mortal Kombat: New Perspectives on Gender and Gaming". Boston: MIT Press, 2008.
[35] Jansen, J., Berden, H., Vleuten, C., Grol, R., Rethans, J., & Verhoeff, C. "Evaluation of cardiopulmonary resuscitation skills of general practitioners using different scoring methods". Resuscitation, 34, 35-41, 1997. https://doi.org/10.1016/S0300-9572(96)01028-3
[36] Castillo, J., Gomar, C., Higueras, E., & Gallart, A. "Checklist-based scores overestimate competence in CPR compared with recording strips of manikins in BLS courses." Resuscitation, 114, e17, 2017. https://doi.org/10.1016/j.resuscitation.2017.02.024
[37] Pulakos, E. D., White, L. A., Oppler, S. H., & Borman, W. C. "Examination of race and sex effects on performance ratings". Journal of Applied Psychology, 74(5), 770-780, 1989. https://doi.org/10.1037/0021-9010.74.5.770
[38] Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. "Digital games, design, and learning - A systematic review and meta-analysis". Review of Educational Research, 86(1), 89-122, 2016. https://doi.org/10.3102/0034654315582065