# Mobile Multimodal Serious Games Analytics

Shoukry, Laila

(2020)

DOI (TUprints): https://doi.org/10.25534/tuprints-00013539



Publikationstyp: Ph.D. Thesis

Fachbereich: 18 Department of Electrical Engineering and Information Technology

Quelle des Originals: https://tuprints.ulb.tu-darmstadt.de/13539

TECHNISCHE UNIVERSITÄT DARMSTADT

# Mobile Multimodal Serious Games Analytics

The Impact of Using Mobile Multimodal Data on Serious Games Evaluation

**Vom Fachbereich**
**Elektrotechnik und Informationstechnik**
**der Technischen Universität Darmstadt**

zur Erlangung des Grades eines
Doktor-Ingenieurs (Dr.-Ing.)

**Dissertation**

**von Laila Shoukry, M.Sc.**
Geboren am 23. Januar 1986 in Kairo, Ägypten

Referent: Prof. Dr.-Ing. Ralf Steinmetz
Referent: PD Dr.-Ing. Stefan Göbel
Referentin: Prof. Dr. Regina Bruder

Darmstadt, 2020
Hochschulkennziffer D17

Mobile Multimodal Serious Games Analytics
The Impact of Using Mobile Multimodal Data on Serious Games Evaluation

Vorgelegte Dissertation von Laila Shoukry, M.Sc.

Referent: Prof. Dr.-Ing. Ralf Steinmetz
Referent: PD Dr.-Ing. Stefan Göbel
Referentin: Prof. Dr. Regina Bruder

Tag der Einreichung: 17. September 2019
Tag der Disputation: 18. Dezember 2019

Darmstadt, Technische Universität Darmstadt

Bitte zitieren Sie dieses Dokument als:
URN: urn:nbn:de:tuda-tuprints-135399
URL: http://tuprints.ulb.tu-darmstadt.de/13539

Dieses Dokument wird bereitgestellt von tuprints,
E-Publishing-Service der TU Darmstadt
http://tuprints.ulb.tu-darmstadt.de
tuprints@ulb.tu-darmstadt.de

## Erklärung zur Dissertation

Hiermit versichere ich, die vorliegende Dissertation ohne Hilfe Dritter nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die aus Quellen entnommen wurden, sind als solche kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Darmstadt, den 7. September 2019

(Laila Shoukry)

## Zusammenfassung

Lernspiele sind eine bekannte Art von Serious Games, die für Bildungszwecke entwickelt werden. Ein Ziel bei der Bewertung ihres Nutzens besteht darin, sicherzustellen, dass sie diese Lernzwecke erfüllen. Geeignete Bewertungsmaße werden in Forschungsbereichen wie Learning Analytics und Technology Enhanced Learning definiert. Im Gegensatz zu anderen Lernmethoden ist das Erreichen von Lernzielen jedoch nicht das einzige Erfolgsmaß für Serious Games. Vielmehr sollen Serious Games wie auch Unterhaltungsspiele Spaß machen und die Anwender motivieren. Dieses Thema wird in Forschungsbereichen wie Game User Research und Game Analytics adressiert.

In dieser Arbeit wird untersucht, wie die Nutzung neuer Datenerfassungsmöglichkeiten durch Smartphones der Forschung auf dem Gebiet der Evaluierung von Serious Games zugute kommen kann. Unauffällige, integrierte Sensoren werden zur Beobachtung von Benutzerinteraktionen in natürlichen Kontexten verwendet. Dies kann dazu beitragen, ein Serious Game aus mehreren Perspektiven und mit tieferen Einsichten zu bewerten. Angesichts des Fehlens theoretischer Rahmenbedingungen für diesen Forschungsbereich werden in dieser Arbeit Grundlagen für die Erfassung und Verwendung von Daten aus verschiedenen Modalitäten für die Bewertung von Serious Games gelegt.

Zunächst wird ein Modell erstellt, das die verschiedenen Bewertungsaspekte von einem Serious Game und deren Zusammenhänge beschreibt. Basierend auf einer Studie von verfügbaren theoretischen Frameworks werden die vier Dimensionen Lernen, Spielen, Usability und Kontext klassifiziert. Zusätzlich werden mögliche Parameter definiert, die sich auf diese Dimensionen beziehen und während der Auswertung beobachtet werden können. Diese werden insbesondere benötigt, wenn herkömmliche Protokollierungstechniken nicht ausreichen und zusätzliche Datenmodalitäten von Smartphone-Sensoren erforderlich sind. Basierend auf einer Anforderungsanalyse wird eine multimodale Evaluierungsplattform für Serious Games, die eine mobile Datenerfassung unterstützt, als Proof of Concept implementiert. Herausforderungen bei der multimodalen Synchronisation, der Verknüpfung quantitativer und qualitativer Daten sowie der Qualität von Daten, die in mobilen Umgebungen aufgenommen werden, werden diskutiert. Mögliche Lösungen werden bereitgestellt. Schließlich werden Studien durchgeführt, in denen diese Plattform bewertet und zur Evaluation von Lernspielen verwendet wird, um das Potenzial solcher Techniken für die Evaluation von Serious Games aufzuzeigen.

## Abstract

Educational games are one prominent type of Serious Games that are designed for educational purposes. One goal of evaluating their benefits is ensuring that they meet these objectives. Some measures for this purpose are defined in fields like Learning Analytics and Technology-enhanced Learning. However, unlike other educational methods, meeting learning goals is not the only success measure of Serious Games. Contrary, complex requirements arise from their nature as games which are discussed in fields like Game User Research and Game Analytics and need to be considered.

This thesis investigates how the use of new data collection opportunities enabled by smartphones as an interaction platform can benefit research in the field of Serious Games evaluation. Unobtrusive integrated sensors are used for observing user interactions in natural contexts. This can be vital for capturing factors that can help evaluate a Serious Game from multiple perspectives and with deeper insights. With the lack of theoretical frameworks covering this area of research, a need for establishing foundations regarding the capturing and use of data from different modalities for Serious Games evaluation was found.

First, a model describing the different evaluation aspects of Serious Games and their relations is laid down. Based on a study of available frameworks, it classifies the four dimensions learning, gaming, using and context. Additionally, sets of feature states related to these dimensions, observable during evaluation, are defined. These are needed to examine aspects when traditional logging techniques are insufficient, and when additional data modalities from smartphone sensors are required. Based on a requirement analysis, a multimodal evaluation platform for Serious Games that supports remote mobile data collection is implemented as a proof of concept. Challenges in multimodal synchronization, linking of quantitative and qualitative data, and quality of data obtained in mobile settings are discussed. Possible solutions are provided. Finally, studies evaluating this platform and using it to evaluate educational games are carried out and demonstrate the potential of such techniques for Serious Games evaluation.

## Contents

# 1 Introduction

Educators are increasingly recognizing the potential of using digital games to engage students in interactive learning environments. Multimedia elements, storytelling, challenge, competition, random access, parallel processing, immediate reward and low level of threat are all features which make many people prefer them to traditional learning media [212]. In general, games that have a certain positive goal beside entertainment (education, health promotion, etc.) are referred to as *Serious Games*. Game-based learning is becoming a more and more acknowledged learning field and the use of digital educational games is on the rise. Some learning game revenues even outperform pure entertainment games [80].

However, it is difficult to predict or reproduce the success of an educational game without having well-established methods for its design assessment. Evaluation plays an important role in the field of digital game-based learning: The decision for a time-consuming and expensive design process instead of traditional media needs to be based on an expected increase in learner's engagement and long-term motivation. It is true that the effectiveness of achieving a learning outcome is the major evaluation goal of any learning platform. However, specifically for educational games, the fun gaming aspect as well as the user experience play an equally important role in design assessment [7, 240]. And when designed for younger ages, edugames should be able to compete with other available games which children would rather play for pure entertainment.

Designing and evaluating a fun and motivating learning game is not an easy task [173]. There is a lack of consensus regarding the best approach to combining gaming and educational theories as well as methods to fully capture and analyze the rich and deep interaction experience with this interdisciplinary software. This makes it challenging to link certain design elements to its success or failure to improve it in further development cycles [117]. Traditional evaluation techniques of non-gaming learning environments or non-educational games do not provide tailored methods which can be directly applied to evaluate the aspects which distinguish learning games from both product categories [142].

The need for more insight into the serious gameplay experience is further intensified with the emergence of novel interaction paradigms which make classical evaluation methods insufficient [171, 241, 35]. Evaluating the whole experience from different dimensions is difficult when relying only on logging mechanisms or traditional testing sessions, due to the multimodal, fast and highly interactive nature of gameplay interactions [171, 241, 35]. This is why more and more studies are using multimodal data (video, audio, screen capture, physiological sensors, mobile sensors, ..) for evaluation [246].

Capturing multimodal data unobtrusively in a naturalistic setting is essential for getting 'real' interaction, affect and context data. As this is difficult to achieve in traditional lab settings, many tools supporting this process are used for testing sessions, especially remote testing of mobile applications. To develop such a system for Serious Games, it is important to establish theoretical foundations for multimodal evaluation covering the dimensions of learning, gaming, usability and context.

## 1.1 Motivation

Affect and cognition have been found to be strongly connected to learner engagement and learning outcomes [31, 113, 202, 213, 168, 214, 222, 89] and exploring them is getting more and more attention in Serious Games research [185, 221]. Examining recorded in-game log events alone cannot fully convey how a player felt or what a learner was thinking. In several studies, it was pointed out that there is a need for multimodal data to fully capture in-game interactions for the evaluation of Serious Games and get accurate results [15, 7].

However, when comparing educational games evaluation studies using in-game data collection for evaluation, it was concluded in [142] that a relatively small percentage of these studies used multimodal data. The authors assumed that this can be due to the costly and time-intensive process which can be

difficult for small research teams to carry out. Not only recording but also understanding and establishing relations across different data streams for the evaluation of Serious Games is a topic which is not fully researched yet.

Another point also raised in the results of the aforementioned study [142] is that most observation data collected during gameplay are not collected in the field but in a controlled environment such as a computer laboratory.Serious Games are played on various devices, platforms and contexts. With the rise in the use of mobile educational games [1] where play is often interrupted with external stimuli, the need for mobile and naturalistic evaluation settings increases as isolated laboratory experiments do not offer the same conditions [118, 8]. Not only does laboratory testing offer an artificial testing setting, but also the obtrusive, difficult to set up and expensive hardware required makes it not suitable for serious games research teams. In addition, the number of playtesters will be limited as they are required to be co-located.

This is why remote testing is gaining more and more popularity in the field of mobile application testing. The user can perform the test comfortably on their own devices and in natural settings and data is sent (even asynchronously) to the researcher who might not be co-located, with no additional hardware cost or expertise needed. Not only does this help in increasing the number and heterogenity of testers, but the different capabilities of smartphones can also be used to provide rich contextual data. Adding multimodal context information will provide valuable insight for Serious Games evaluation by helping researchers understand why an event occurred during the process of playing and learning [141, 88]. As known from the field of multimodal recognition, adding more modalities helps in disambeguating the results of other modalities and, in the current case, the results of examining text logs. Without this additional data, some interesting behaviors and intentions of the user can get masked when just aggregating log statistics. For instance, when classifying student dialogue acts, adding gesture and posture information were found to be useful and helped researchers recognize the intentions of utterances [72].

Another advantage of using more than one modality for observations is that this might help detecting interesting correlations between different modalities which support or negate certain research claims.

The main motivation for this research was thus the need for (mobile) multimodal frameworks for the evaluation of Serious Games and the chance of exploiting the new technological landscape for capturing valuable information about a Serious Games session unobtrusively. While there are a number of multimodal evaluation platforms available for testing interactive systems in general (see a comparison of features in Chapter 2), none of them is especially tailored for testing mobile educational games.

During the initial phases of this research, it was found that the absence of comprehensive theoretical frameworks and concepts tackling the multimdoal evaluation of Serious Games, especially on mobile devices and thus in mobile settings, made even the first step difficult for many researchers. Although this is increasingly being researched in other disciplines, there is surprisingly almost no research on this in the field of serious games. This is an attempt to compensate this gap and help Serious Games research advance and adapt to the changing technology and research landscape.

## 1.2 Research Challenges and Goal

Although remote testing in the field offers many chances for Serious Games evaluation which motivated this research, tests in real environments suffer from dynamic factors affecting recorded data quality such as variations of illumination, exposure, orientation, distance, stability and noise. Such factors are more difficult to control than in laboratory settings and result in missing or noisy data [35]. Thus this is one challenge which has to be addressed by the current research.

Even the advantage of scalability and an increased participation in remote testing comes with its own limitations. Having more testers means having more data, which can be desirable when applying statistical methods but maybe difficult to analyze when doing so manually. Even if applied only on a smaller scale in earlier design stages, making sense of data needs observing a session from multiple modalities at the same time [96]. Thus to make use of captured multimodal data, researchers need tools

to facilitate this process by pre-processing and synchronizing data, making it ready and easy to navigate. This will be a very important requirement in the current work.

Another important issue worth considering is concerned with the data analysis phase. After capturing and pre-processing the data, will researchers more probably use automatic or manual ways to make sense of and draw conclusions from the data? Although there is lots of research on using fully automated methods in affect recognition, it is still found to be very challenging. This is due to the heterogenity of users, expressions, contexts [185] and for the current case also the heterogenity of serious games, interaction paradigms and design goals making it difficult to obtain accurate and generalized results [221]. In fact, even among close populations, small age differences were found to have considerable effects on test results. For instance, one study [89] showed how facial expressions related to learning experiences and self-efficacy differed between middle school and college students. Thus automatic recognition, despite eliminating the need and risk of manual annotation, cannot yet be relied upon alone when it comes to using multimodal data for evaluation. Observation by an expert, if done in an effective way and using efficient analytics tools, can yield more reliable and applicable results, especially spotting significant behaviors in a field where there is still a lack of standard corpora or generalizable methods and definitions. Serious Game researchers and developers (also teachers), with limited time and budget for their evaluation and who need a theoretical framework and user-friendly platform for multimodal evaluation, will thus be the main target audience for the current research. Nevertheless, the use of automatic analysis modules should be easy to integrate into the framework and used on captured and pre-processed data. In addition, the platform can be used to capture a large dataset for training different detection systems.

As an added value, applying multimodal evaluation techniques on a large set of heterogeneous games in different settings and with different users can yield aggregated results which can help create a personalized content- and context-based Serious Games retrieval or recommender system, which is another motivation and application field for the current research.

Research focus is *not* on the following aspects as they are out of the scope of this work:

- affect recognition and its accuracy
- networking, off-loading etc.
- (cost-)effectiveness of the applied games from the perspective of game development, benefits and costs for educational institutions
- feedback loop for Serious Game design (needs subject matter experts)
- long-term learning evaluation (enhancing performance or loosing interest over time), transfer of learning, etc., as the focus is only on a certain playtesting session

## 1.3 Context

The start of this research direction was influenced by the Serious Games authoring environment StoryTec [166] which was built by the Digital Storytelling group at the Center for Computer Graphics in Darmstadt and further cultivated by the Serious Games group at the Multimedia Communications Lab (KOM) at the Technical University of Darmstadt. StoryTec is built upon an internal model considering updates in the learner and player model during play as well as the storyline based on the Narrative Game-Based Learning Object (NGLOB) Model [81, 82], another earlier contribution of the same research group. In addition to authoring, considerable research is done on the monitoring and evaluation of Serious Games experience, be it in games for health or education. A rapid prototyping tool for the StoryTec authoring environment displaying updates in the internal models was one of the group's earlier attempts in helping researchers evaluate educational games created with StoryTec [217]. Continuing to adapt to recent research needs, it is the goal of the current research to investigate ways in which more multimodal data, in addition to traditional text-logs, can enrich the Serious Games evaluation process by capturing more elements of the play-learning experience. Hence not only do the theoretical models of this research contribute to the body of research at KOM and address an identified research need, but, through the

prototypical implementation, StoryPlay is extended by multimodal features, encouraging more applied research in this field. The requirements for this work arise from a real need identified by the group in various research projects and not on mere speculations.

## 1.4 Research Questions

Based on the needs identified in the Serious Games Analytics research field, the main goal of this research is to investigate the impact of using (mobile) multimodal data on Serious Games evaluation. This leads to the following research questions:

**Research Question 1:**
Which parameters constitute a conceptual research model for Serious Game evaluation?

**Research Question 2:**
Why and when do we need to add multimodal data to interpret log events in Serious Game evaluation?

**Research Question 3:**
Which information do we need from multimodal data for Serious Game evaluation and how can it be captured?

**Research Question 4:**
How can this information be linked to recorded log events for Serious Game evaluation and what are the associated challenges?

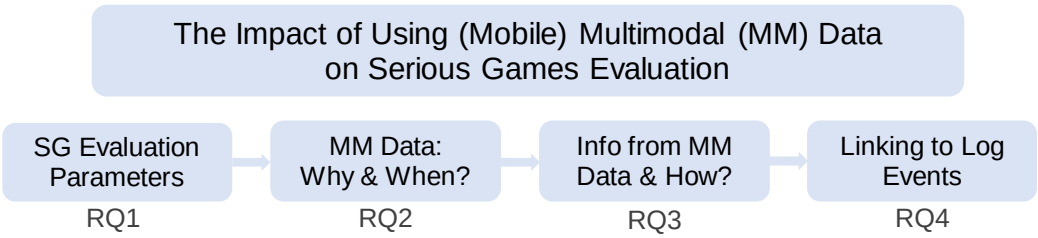



**Figure 1.1:** Research Goal and Questions.

## 1.5 Contributions

The initial research analysis carried out revealed a need for defining concepts and mechanisms for non-invasive multimodal capturing in natural settings to support Serious Game evaluation. Thus as a first step, a theoretical framework for multimodal Serious Games evaluation based on a review of relevant research is proposed. The *LeGUC framework* (Learning, Gaming, Using and Context) describes the different aspects of a Serious Game which are to be evaluated and classifying the four dimensions: Learning, Gaming, Using and Context. Under each of these dimensions, different requirements need to be met to increase the quality of a Serious Game. These evaluation aspects are further categorized into features that can be evaluated without real users and others which need testing sessions with participants to be investigated, as they depend on traits, abilities and preferences of different users. These include aspects which need capturing and analyzing affective and cognitive states as well as context parameters. *The Reasons and Responses Model* demonstrates when and why more multimodal data is needed for Serious

Game evaluation. Finally, as a proof-of-concept, a platform for supporting remote and asynchronous replay and analysis of recorded Serious Game testing sessions was developed based on the outlined conceptual framework. Parts of this thesis have been published in papers which are listed in Appendix A.

| | | | |
|---|---|---|---|
| Research Analysis | Serious Games Analytics | SG Evaluation Models | MM Evaluation Platforms |
| Identify Gap | Novel Settings and Modalities | Covering All Dimensions | Tailoring to SG Evaluation |
| Concept | LeGUC Model | Reasons & Responses Model | LeGUC States |
| Implementation | Requirements and Mockup | Implementing Modules | Overcoming Challenges |
| Evaluation | MM Benefits: What & When | Sensor-based Video Quality | Platform Evaluation |

**Figure 1.2:** Research Contributions.

## 1.6 Thesis Structure

The remainder of this thesis is structured as follows:

**Chapter Two: Foundations**

A foundation for the next chapters is first laid out explaining research terms and fields directly related to our topic. Areas like Learning Analytics, Educational Data Mining, Game Analytics, Serious Games Analytics, Affective Computing and their intersections and differences are presented. The different steps of Serious Games Evaluation are discussed. This will help establish understanding paving the way for the following chapters.

**Chapter Three: Related Work**

In this chapter, a review of related research is presented. First, a survey of user evaluation studies using multimodal data then investigates why and in which areas and contexts these methods are used. Also, the topic multimodal data capturing on mobile phones is researched and different related studies in this field are discussed. Next, different theoretical frameworks on Serious Games Evaluation are compared to determine which dimensions they cover. Finally, available evaluation tools and platforms are then presented to help in gathering requirements for creating such a platform tailored for Serious Games.

**Chapter Four: LeGUC Serious Games Analytics Model**

Based on the related research in Chapter Three, the LeGUC Serious Games Analytics Model is presented for Serious Games Evluation as an attempt to fill the identified research gap and as a foundation for a Multimodal theoretical framework. After determining what is to be evaluated in a Serious Game, we need to determine how these aspects will be measured and where multimodal data can add more value to the evaluation. So the LeGUC Model serves as a foundation for the more in-depth model aimed at multimodal data which will is presented in this Chapter: The Reasons and Responses Model. These

theroetical frameworks are then applied on a case study to explain them on real data. Finally, challenges of using multimodal data are discussed to suggest requirements for this project.

**Chapter Five: StoryPlay Multimodal: Mobile Multimodal Serious Games Analytics Platform**

In this chapter the design requirement for a Mobile Multimodal Serious Games Analytics Platform are presented based on the previous chapters. This is used as a foundation for the design and implementation for StoryPlay Multimodal platform. Then, Implementation aspects of our platform are presented as well as challenges and solutions.

**Chapter Six: Evaluation**

The evaluation process and results of the implemented platform are presented in this chapter as well as an application with different serious games and testers.

**Chapter Seven: Conclusions and Future Work**

The work is concluded in this chapter, answering our research questions and giving an outlook on future research which can benefit from our results and analysis.

## 2 Foundations

Serious Games apply concepts of game technology in a broad application spectrum ranging from educational, training and simulation to persuasive and social impact games or games for health. In this thesis the term Serious Games is used to refer to digital educational games, although this is one application area of Serious Games. Because of its interdisciplinarity, the evaluation of Serious Games lies in the intersection of and benefits from metrics and strategies from several different fields. Hence, it is important to review some important fields which are considerably related to this topic. This chapter gives an overview over relevant areas and terminology which will be referred to in the rest of the thesis. A main problem in providing such an overview is the variation in terminology, research background and objectives among different researchers when discussing similar topics. This can be facilitated by starting with locating this interdisciplinary research area and related terminology in the research landscape. Figure 2.1 shows an overview of research areas to be discussed in this chapter and their intersections. The focus of the current research, the topic of multimodal evaluation of Serious Games, lies in the intersection of all four discussed research fields. It can be called "Mobile Multimodal Serious Games Analytics".

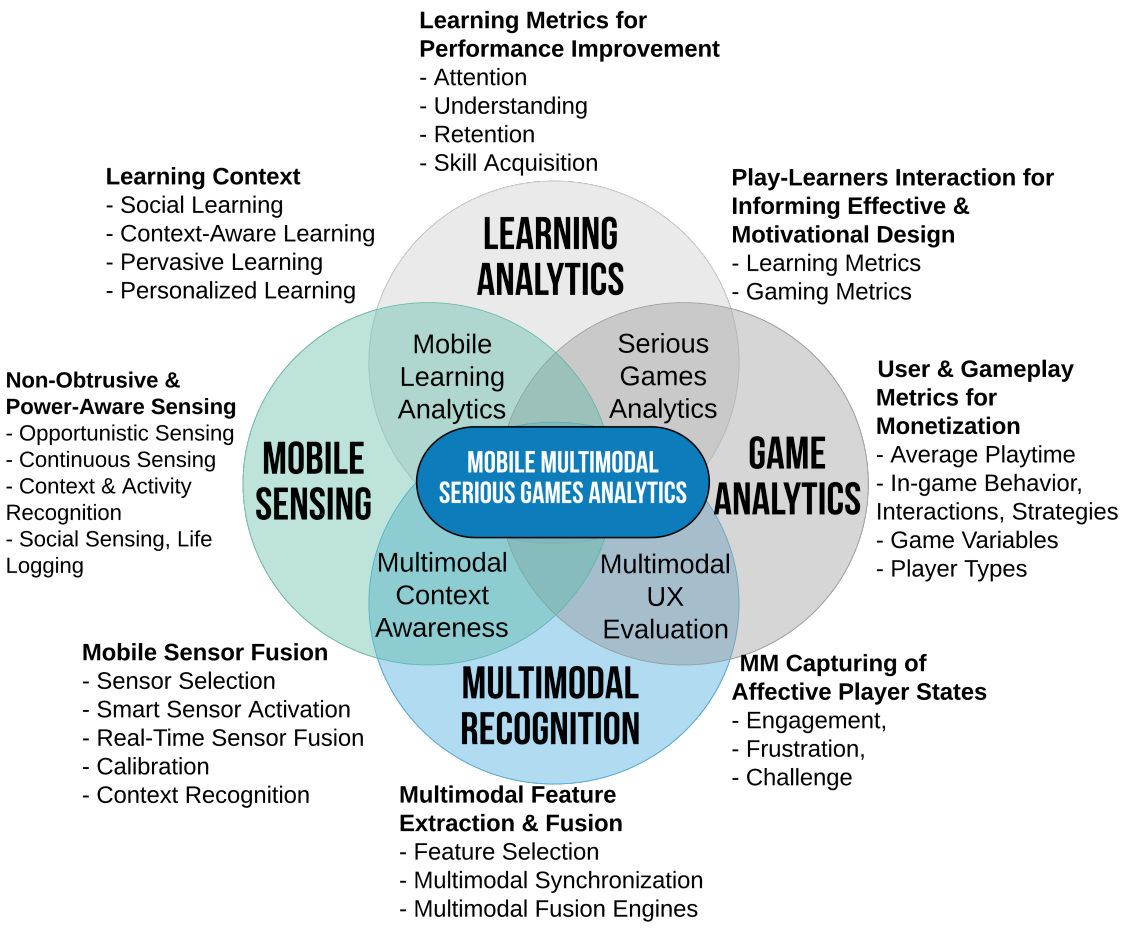



**Figure 2.1:** Topics and Intersections of the Discussed Research Areas.

## 2.1 Learning Analytics

With the digital evolution in education comes a necessity for improving assessment strategies accordingly which are integral to any learning process. As educational data volumes increase and learning practices get more and more diverse, finding appropriate evaluation techniques becomes a challenging task. Although technological advances also enable advanced tools for gathering and analyzing data, research on how to best utilize this for enhancing learning and assessment in novel educational settings still has many open questions [17].

The emerging fields of Learning Analytics (LA) and Educational Data Mining (EDM) attempt to address these issues by exploring models and techniques for making efficient and effective use of educational data. This includes capturing, tracking, aggregating, analyzing, and visualizing/utilizing information about learners' interactions with learning content and their learning progress. Both disciplines have the common goal of using these traces to improve learning but their methodologies for exploring and using data are slightly different. While EDM is more concerned with analyzing and feeding data into automatic adaptation, prediction and recommendation algorithms, LA directs more emphasis to appropriately visualizing these mechanisms for the user (instructor, instructional designer, institution, parent and/or learner) and empowering him/her to interpret and intervene in the process [249]. Its learning benefits have been manifested in many studies where they were implemented in different contexts [147, 251, 9, 60].

The field of Learning Analytics describes the "measurement, collection, analysis and reporting of data about learners and their contexts" to improve learning [250]. It establishes conceptual foundations as well as application-specific recommendations for making efficient and effective use of "information about learners' interactions with learning content and their learning progress". Metrics typically extracted in Learning Analytics include information about the learners (independent of their interaction with the learning environment) in addition to context-specific data extracted during interaction [142].

As the learning aspect is what distinguishes Serious Games (here digital educational games) from regular games, the evaluation of Serious Games uses methods from the field of Learning Analytics. The assessment of learning in Serious Games, as in any other learning environment, can take place before, during and/or after the interaction with the game. Assessment of learning before using the game can inform the design of the game, help personalizing the game or serve as a pre-test for performance comparison. Assessment of learning during the game can be used to adapt the game on-line, to monitor learning progress by teachers or institutions, as self-assessment for learners to reflect on their learning and understanding, as part of the learning process to improve their memory recall or as a means to evaluate the design and effectiveness of a game level or module. Assessment of learning after finishing a game helps evaluate the effectiveness of the game, help learners know what they need to improve or, if used (merely) as an assessment tool, to assist instructors/institutions to assess learner achievements against intended outcomes.

Thus, for the evaluation of Serious Games, Learning Analytics provide an objective and cost-effective approach for justifying the expense of using them in learning contexts. Hereby, the critical question is how to best collect academically meaningful information from learners without disrupting their gameflow. In addition, interacting with a Serious Game is usually more complex and produces larger datasets than interacting with a regular learning environment. In this regard, the field of Game Analytics and, more so, Serious Games Analytics provide suitable solutions.

## 2.2 Game Analytics

While Learning Analytics focuses on learning-relevant information, Game Analytics, as the term implies, focuses on gameplay interactions and gives enjoyment of the game experience the highest priority [142]. Metrics used in Game Analytics are also typically more profit-oriented and are used to inform business-intelligent decision-making and improve game development cycles [142]. Monitoring the effect of design

updates on players is crucial for continuously improving and thus increasing revenue. These insights into real and dynamic data in huge amounts (also called game telemetry) has been made possible with the digital advances and thus the field of Game Analytics is relatively young. The term Game User Research also describes a research field investigating the interaction between players and games [228]. In addition to general business data, Game Analytics trace players' actions inside the gaming environment in order to gain a better understanding of their preferences, detect crucial design flaws as well as identify best-gamers. However, although games offer rich information about players' progress, the game industry is not focusing on assessing competencies gained through the game as the games are not designed with this purpose [142]. Thus, pure Game Analytics cannot be directly applied to Serious Games which were not created for mere entertainment.

**Telemetry**

The term telemetry has been traditionally used in association with data collection in different fields and is applied for remote mobile app data collection to a server [141]. In the field of game analytics, this technology is used for (educational) games so that players' interactions can be remotely traced for analysis [108, 139].

## 2.3 Serious Game Analytics and the Serious Game Evaluation Process

It would appear reasonable that combining metrics from the fields of Learning Analytics and Game Analytics would enable an effective way of evaluating Serious Games considering both their learning goals and their highly interactive nature. However, some metrics of the two fields are actually conflicting. For instance, failure associated with exploration inside a gaming environment contributes to the flow needed to stay captivated inside the game whereas failure is not typically a desired experience inside a learning environment and subsequently can be treated as a conceptual error. Another example is the time to complete a certain task: While in games it is desired to make fast decisions and complete tasks in a short time, learning needs more time to allow reflecting and critical thinking. Users of Serious Games are players as well as learners, are thus referred to as play-learners [142] and need special metrics to measure their performance as well as evaluate the quality of their playing experience. This is where the field of Serious Games Analytics comes into play, tracing play-learners actions and behaviors and assessing their performance with the purpose of improving Serious Game Design. Thus, to be able to apply Serious Games Analytics, the broader research topic of Serious Games Evaluation needs first to be explored.

### 2.3.1 Serious Games Evaluation

#### Primary Objective of Evaluation: Why to measure?

When reviewing research dealing with assessment in Serious Games, it becomes evident that, depending on the primary purpose and context of the studies, they are either directed towards assessing the learner or assessing the Serious Game. Assessing the learner online can help in adapting the game or making predictions. This is also referred to as in-game analysis [97]. Assessing the learner can also be useful, together with purchasing habits, for making recommendations. Integrated assessment in Serious Games, also referred to as "Stealth Assessment" [247], can serve teachers, institutions or the learner himself for monitoring learning. Using games as assessment tools in academic contexts is becoming more and more popular due to their interactive, non-invasive and engaging nature and the big amount of gameplay data they can offer [134]. Assessing the Serious Game is typically done off-line, also referred to as posterior analysis [97] and has the aim of proving the fun and effectiveness of the game or uncovering possible shortcomings to be fed into further development cycles or new games. As there is no common agreement

on the best method to evaluate and improve the design of Serious Games, this topic will be addressed in more detail in the next chapters.

### Aspects to Evaluate: What to measure?

By inspecting the name "Serious (or, in this case more precisely, Educational) Games", one can already derive two main dimensions of parameters which need to be considered in Serious Games evaluation: Learning and Gaming. However, by reviewing literature on Serious Games evaluation, a third dimension was found to also play an important role, even more for Serious Games than for other software categories: the dimension of Usability and User Experience. This is because usability issues can greatly hinder interactions with serious games, which directly has a negative effect on both learning and gaming, and may thus make all design and development efforts on the other two dimensions go in vain [207, 28, 192, 175, 205]. As serious games development teams typically have a much lower budget than the entertainment game industry, not much is usually left for ensuring good usability which can explain the mostly inferior quality of experience. In addition, many users of Serious Games are not regular gamers and may thus need special measures concerning usability. The main difference to regular usability testing in Serious Games evaluation is that some obstacles to proceeding introduced in a game are necessary components for adding challenge while these could be considered as negative issues in a normal usability setting.

### Evaluation Methods: How to measure it?

After determining what aspects about the serious game are required to be evaluated, it is time to decide for the appropriate evaluation procedure. For that, the evaluators, the time and place of the evaluation as well as the tools used are to be planned. Evaluators of a serious game can be real users or surrogate users who perform expert evaluation [193]. These can be educational, game design, interaction design and/or content experts. According to the time and place of the evaluation, different types of mechanisms can be differentiated. Ex-situ evaluation refers to remote data collection without the need for co-location [142]. In this case telemetry refers to remotely tracing the in-situ generated gameplay data from a gaming environment [242]. In synchronous evaluation, the users and their actions are traced/analyzed in real-time whereas in asynchronous evaluation, gameplay data is reviewed any time after the evaluation session. Laboratory-based evaluation needs testers to be co-located which limits participation and the testing environment is controlled but artificial. Remote or mobile evaluation gives the possibility of employing testers from all over the world in their natural environments, without the need for special, expensive and obtrusive equipments but has the downside of having an uncontrolled and dynamic environment and may suffer hardware limitations. Thus, if the serious game to be tested is mobile and needs to be tested in a naturalistic setting, a remote evaluation may be appropriate. [211] reviews different elicitation methods for studying learners' affect in intelligent tutoring systems. Similarly, [83] discusses different methods and modalities used in the evaluation of serious games. Static (user-independent) product aspects can be evaluated just by inspection or expert heuristic evaluation without the need for real users to do the testing [245]. With real users, qualitative or quantitative methods can be used. Qualitative or manual evaluation methods include observation, think-alouds, pre-post-tests, self-reports, video analysis or surveys based on heuristic rules. Quantitative methods usually rely on game telemetry or automatic capture and analysis of multimodal data.

### Data: What will be captured to measure it?

According to the aspects to be evaluated and the evaluation methods applied, the data to be captured will differ. Qualitative studies will result in questionnaire/interview responses, observation notes, think-

aloud protocols and/or screen/video/audio recordings. Data from quantitative studies will most probably include log files but can also include video and audio recordings which are commonly used for detecting visual and audio features and nowadays also frequently rely on physiological and/or mobile sensing. For recording in-game interactions, log files are mostly used, while for detecting affective and cognitive states, visual and physiological features are mostly extracted, although many studies also attempt to do this using pure logging alone. For context, mobile sensors are usually used. In many evaluation studies, data from multiple modalities is used.

#### Analysis: How to make sense of data?

According to evaluation method and data captured, different analysis methods can be used. Qualitative data can be explored and annotated by experts manually to make conclusions [240]. Quantitative or automated evaluation methods can use data mining or unsupervised machine learning on logging and multimodal data. Hybrid methods use supervised machine learning or fuses self-reports with logging.

Analytics are often seen to be a relatively objective and cost-effective evaluation approach compared to relying solely on self-reports and pre-/post-tests. It also has the ability to give real-time insight into possible shortcomings of an educational game. In addition to assessing games, the interactions of players during gameplay sessions can provide a lot of data which can be used for identifying play-learner attributes, strengths and weaknesses as well as measuring learning progress where they can even outperform traditional tests.

Integrated into Serious Games, analytics could thus perfectly fit as a tool providing implicit insight into a learner's knowledge state within the game. The wide variety of educational games poses a significant challenge for defining a general methodology for integrating learning analytics into serious games in an effective way [230]. We will therefore in the next section present several suggested solutions to this problem covering different aspects like modeling decisions, data collection, reduction, aggregation and analysis as well as effective use of results.

Figure 2.2 summarizes steps in planning and conducting serious games evaluation.

### 2.3.2 Serious Games Analytics

#### Modelling for Analytics

To successfully assess learning using learning analytics, the learning domain, the application concept as well as the learner should be modelled in a way which facilitates data extraction and analysis. These models should preferably also be stored in separate files in an easy-to-process format [230]. In this regard existing models like the Competence-Based Knowledge Space Theory (CbKST) are usually used as a foundation [128, 73]. This theoretical framework requires learning domains to be modelled as a prerequisite competency structure to make the process of inferring knowledge states more efficient. Open Learner Model (OLM) [40] is becoming a popular term among LA researchers as it requires presenting to the learner an understandable visualization of his/her current knowledge state. Several studies have shown how the OLM approach improves learning outcomes [288, 174]. In some studies, the learner model was even made directly editable by learners. Using the *Mobile Open Leaner Model* the learner can also "carry" his/her learner model with him/her and exchange it with other students to facilitate peer tutoring [39]. Adaptive learning games, like adaptive tutoring systems, already need to employ a learner model for their adaptation mechanism. One suitable model for this kind of games which considers not only the knowledge state but also the player type and the narrative aspects of the game is the concept of Narrative Game-Based Learning Objects (NGLOB) [82]. Here the representation of context information consists of a triple vector for each scene containing information about the narrative context (how appropriate a scene is for which step in which story model), the gaming context (how

appropriate a gaming situation is for which player type) as well as the learning context (all associated and prerequisite skills of a learning situation). Dependencies between skills are modeled as a graph based on the CbKST [128]. Before being able to apply such an approach, relevant information should exist and be appropriately structured to provide the following models: A Competency Model, an Evidence Model and a Task/Action Model [248]. The Competency Model should present a fine-grained specification of competencies which should be assessed. The Evidence model should relate different actions and behaviors of a player within a learning game to the different competencies. In this regard, Domain Structure Discovery refers to the problem of mapping tasks to knowledge components or skills [264]. Conceptualizing how a certain gaming context should be best designed so that player actions required will result in measurable inference about their competency level constitutes the Action Model. This approach in designing learning environments is referred to as Evidence-Centered Design (ECD) [172]. As discussed before, in addition to modelling learning, other aspects like engagement are worth modeling in the case of evaluating educational games. Therefore it is suggested in [119] to not only collect evidence

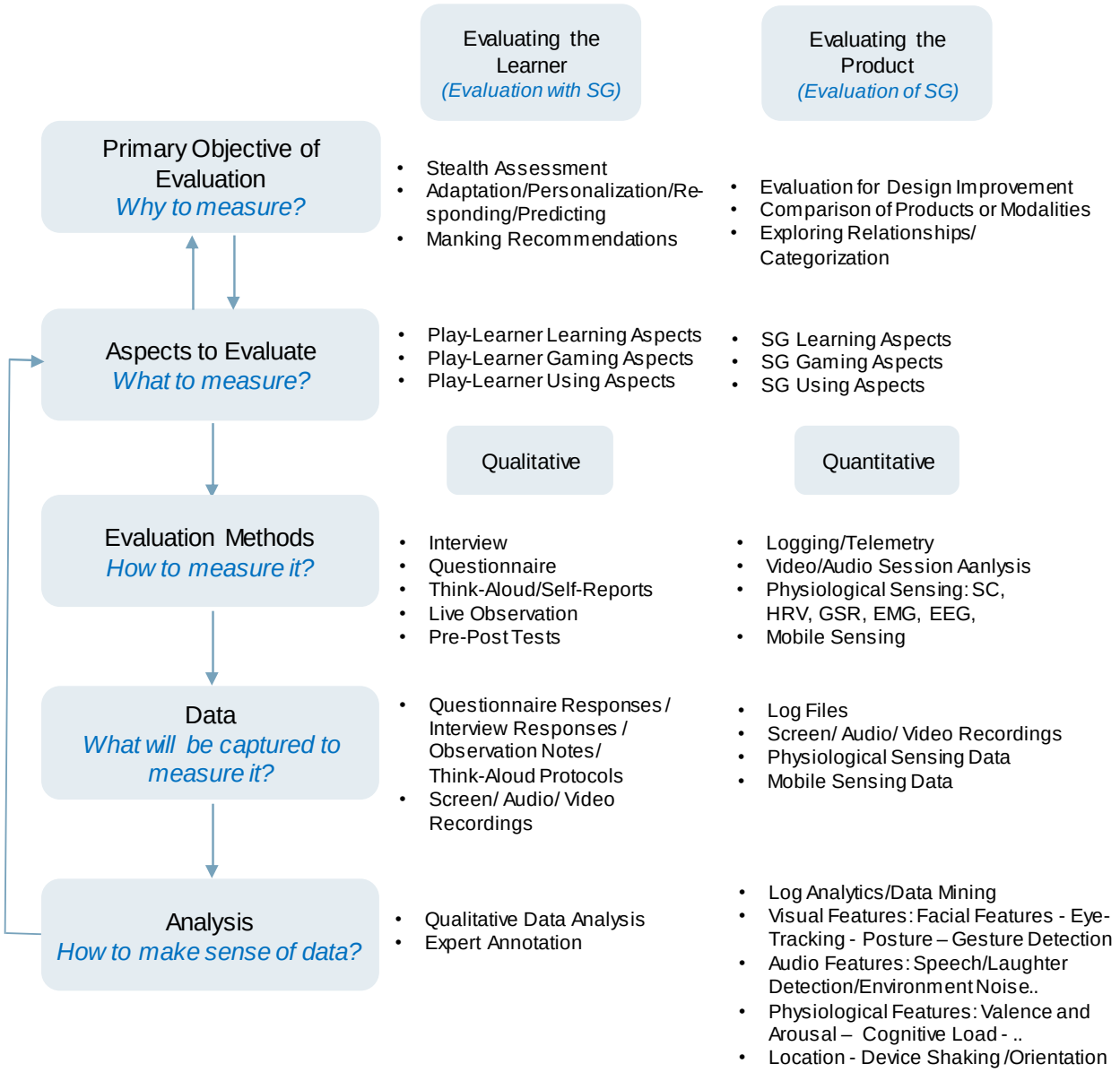



**Figure 2.2:** Steps in Planning and Conducting Serious Games Evaluation.

on the player's state of knowledge using a so-called Skill Assessment Engine (SAE) but also considering his/her motivation level through a Motivation Assessment Engine (MAE) as well.

### Choosing Data

Determining which information needs to be extracted is an essential step for using learning analytics. This depends not only on the learning goals, setting and tasks but also on the game genre, mechanic and platform.

#### Intensive vs. Extensive Data

One possible classification of data to be collected using learning analytics depends on whether quantity or quality is desired. Collecting data from a large number of users with only few information gathered about each user results in extensive data while focusing on a limited number of participants to make deeper and more-detailed observation produces intensive data [209]. Extensive data is primarily for Educational Data Mining where patterns are recognized across large data sets while intensive data performs better at recognizing patterns across multiple different data streams of one and the same user over time. In some cases an approach combining both intensive and extensive data may be best as both complement each other ensuring that no significant patterns are overlooked.

#### Single-Player vs. Multiplayer

Collecting data from multiplayer games poses its own challenges arising from the additional social component. Learners' interactions in a collaborative learning environment, within a Learning Management System (LMS) or a multiplayer Serious Game, are a rich source of data which can be exploited using LA. Data extracted from similar environments can be fed into a social network analysis process to identify aspects of collaborative learning through relations and structures [189]. A diagnostic tool is presented in [126], where social networking principles are used for in-class peer assessment offering teachers a teacher supervision panel allowing monitoring task solutions and filtering collected information according to their diagnostic or instructional interest.

#### Generic vs. Game-Specific Traces

In [230] the authors define a set of generic traces which can be extracted within learning game experiences for learning analytics. The set consists of game traces (starting, quitting and ending a game), phase traces (starting and ending phases), meaningful variable traces (update of variables) and input traces (clicks and key presses). The *Game Start* trace not only contains information about the time the game was started but also basic information identifying the user and describing context and demography through technical data. The *Game End* trace records when the game was finished and, if the game has many endings, which ending was reached. If the game is quit before reaching the end, the *Game Quit* captures the context in which the session was interrupted. Phase changes can also be mapped to storytelling elements and are linked to sub-games or learning chapters in a game. Here, the *Phase Start* and *Phase End* traces can be used to identify when a phase has started/ended and whether or not it was completed with success. Information contained in input traces can be input source, type of action and associated data. Similarly, the traces collected and visualized in StoryPlay [230] are based on the theoretical framework of the NGLOB framework described earlier [82]. In addition to logging all user inputs, active variables, and the time taken for each scene, the system records updates to the internal state in all three adaptation dimensions: storytelling, gaming and learning. This consists of a history of visited scenes with their respective adaptation algorithm parameters based on associated and prerequisite skills, the narrative context as indicated by appropriateness values, player attributes describing the player model and the player's skill state. Generic game logs provide valuable information which can be used to assess learning games and identify strengths and weaknesses in their design. It also makes comparing different games possible. However, not all information is equally meaningful in all

types of games. In addition, games can produce huge amounts of partially irrelevant data which must be filtered for analysis. For a detailed exploration and yet efficient analysis, it is thus recommendable to additionally tailor analytics to the specific game design features and evaluation requirements. Ideally, the game should be designed with analytics in mind. In other words, all major game mechanics in an "analytics-efficient" design are chosen in such a way as to directly reflect a learner's skill or behavior interesting in terms of evaluation [272] . An example of such a game is *Save Patch* presented in [116] where all player actions are tied to mathematical operations. The logging structure for this game was also designed to ensure no key information gets lost and no data significant to evaluation is overlooked. In [50], the authors suggest logging data at the "finest usable grain size", i.e. recording clearly defined, unambiguous events sufficient for understanding the context.

### Capturing Data

To allow the integration of learning analytics, the game platforms should be able to record game traces and are usually also required to allow sending them to external servers for collection [230]. However, data useful for learning analytics is not limited to activity logs created by the game engine. Measuring engagement and learning can be done by combining data from different sources.

### Aggregating Data

After capturing extensive data from different users and different sources, datasets should be merged for statistical processing or data mining to be able to extract information from it. For intensive data, aggregation is also required to combine multiple streams of data captured by different devices about one and the same user. A key challenge for this process is the current lack of interoperability [54].

#### Aggregation across Users

Logs typically contain a large amount of data and thus usually undergo preprocessing before they are ready for analysis. Structuring, segmenting, filtering and normalizing raw data is done according to the application at hand. Using session identifiers, data from different users can then be joined into one central database. To enable this, the log files generated on all machines should be using the same data format. There have been several attempts to standardize xml-based formats of log files for educational data mining applications. In [125], a data format for logging interactions of learners with tutoring systems is presented. Similarly, a format for encoding interactions within Computer Supported Collaborative Learning (CSCL) environments was introduced in [66]. An example of aggregating logs of multiple players in learning games was implemented for the tool StoryPlay described earlier where playtraces from individual sessions are combined into one comprehensive spreadsheet [217]. In [229], the use of *aggregation models* is proposed which use semantic rules to map game actions or states to meaningful expressions under which similar events are grouped.

#### Aggregation across Modalities

When gathering multimodal data to enrich the interpretation of log files, researchers are faced with many challenges during the aggregation phase [32, 66, 133]. Synchronizing data captured in parallel using different devices is necessary for observing behaviors at specific timestamps across data streams and for analyzing situations, confirming claims and drawing conclusions. Tools like *Replayer* are available which enable synchronized play of streams they captured and proved useful in such cases [176, 66]. In [32] the authors describe their time synchronization process for coordinating the encoding of several video streams on different machines. Digital ink files were hand synchronized using the *ChronoViz* multimodal analysis tool with their corresponding audio-visual files. An approach for synchronizing eye-tracking data and log events with EEG signals was presented in [133], where they were imported as hits at significant timestamps of the EEG traces to find correlations across channels.

### Analyzing Data

After the aggregation stage, data can directly be used for reporting, i.e. statistics and abstracted overviews can be provided to the instructor/game designer or the play-learner whose task would be to extract useful patterns from the data. If the system is to analyze data automatically, as the case with adaptive educational games, then this stage is the most complicated, especially in games as real-time processing is required for personalization. Expressing it from the learning perspective, the gathered data should help get inferences about general traits and abilities of the learner, his/her general knowledge state, his/her situation-specific state, his/her learning behaviors and his/her learning outcomes [70]. From the gaming perspective, measures to be derived are general game performance, in-game learning and in-game strategies [50].

In the ECD approach described earlier, the Evidence Model describes the rules which govern the interpretation of in-game sources of evidence to infer competencies. Usually, algorithms are applied during learning sessions to update the competency model of the learner according to achievements and failures exhibited at runtime or more complex heuristics. Here, Bayesian networks can be used to update associated probability distributions of the different competencies [135]. In the 80Days project [119], the modeled knowledge structure is stored as a binary matrix, parsed at design time and then loaded during gameplay into the runtime component of the Skill Assessment Engine. According to the player's game interactions propagated by the Game Engine, the Skill Assessment Engine updates probabilities of corresponding skills whereas the Motivation Assessment Engine updates probabilities of aspects related to the learner's motivation like attention and confidence. In [116], Cluster Analysis was used to identify solution strategies and error patterns of players from learning logs. Error patterns were further classified into errors resulting from mathematical misconceptions and those related to game strategies.

The author of [230] specified rules for interpreting semantic information from his/her gathered generic game traces described earlier. The idea is that by defining conditions on the variables like time spent on tasks or phases, values of in-game-variables or more complicated rules depending on the context, more insight into game experience and learning achievements can be gained. Similarly, it was possible, by focusing on certain intervals of biometric data streams where certain significant events were recorded in log files, to answer semantic questions [133].

Examples are finding out where the attention of the user was directed when a certain event occurred or if certain physiological responses are related to certain interactions. In another study a model based on Markov logic networks was proposed for recognizing player goals through their interaction in non-linear games [92]. Prediction is in this regard also a popular approach for analyzing logs. For example, real-time affect can be predicted in learning games using computational models as described in [132] or [223].

### Deploying Results

Using inferences resulting from the interpretation stage can be done in two ways: Either this information is communicated to the instructor or the learner to empower him/her to make decisions on possible measures or interventions, or they are directly fed into adaptation mechanisms implemented in the game-based learning environment. The system can respond by choosing the appropriate next learning object or narrative event (macro-adaptivity) and/or adjusting aspects within a learning task like task difficulty or feedback type (micro-adaptivity) [119].

An example for making use of logs for enhancing learning experience in learning games through adaptation is Leo's pad, a smartphone application for children offering an early learning curriculum in an interactive playful environment [276]. Along with this application, Parent's Pad provides parents with useful insight into the learning activities and progress of their children. *Zoodles* [1] is another example of

[1] http://www.zoodles.com (last accessed January 2018)

a parent dashboard for mobile educational games. Providing this kind of readable information to guide a learning experience is one of the main goals of LA. This is why a lot of studies were carried out to identify how to best visualize log data and learner models [64]. Hasse Diagrams are a popular way for graphically representing competency structures.

In StoryPlay, visualization of logging information is realized using a replay component as well as different real-time visualizations of narrative structure, player model and skill tree, respectively. In [230] clicks on the screen were aggregated and visualized using heat maps to help identify which objects on the screen received more attention. The learning analytics toolbox eLAT [65] employs several visualization mechanisms which are not specifically targeting games but cover many aspects interesting for visualizing LA in Serious Games like real-time operation, extensibility and interoperability.

Authors of [170] propose a technique for real-time learning analytics visualization for educational games applied on the example of an adventure game teaching computer networks. Two powerful and popular web analytics tools available for free use which can also be incorporated into learning apps and games are *Google Analytics* [2] and *Piwik* [3].

Virtual worlds like *Second Life* [4] and *OpenSimulator* [5] are increasingly being used for learning by creating immersive learning scenarios making use of the natural and rich environments and the multimodal and collaborative interactions as well as logging reports they offer [120]. An example for such a learning scenario is "Chatterdale Mystery", an English learning adventure game developed for OpenSim, is presented by the authors. Along with this game and as part of the Next-Tell project, a teacher control center software was implemented which offers easy-to-use tools for aggregating, analyzing and visualizing log files generated by OpenSim together with other sources within learning adventures allowing teachers to define and integrate their own evidence rules.

Figure 2.3 depicts the process of applying Analytics in Serious Games as described in [240].


**Figure 2.3:** Process of Applying Learning Analytic in Serious Games as Described in [240].

## 2.4 Multimodal Recognition

Today more advanced technologies and sensors can capture data with different modalities. As Serious Games claim to offer learning content in a more "fun" environment, assessing User Experience is a crucial part of Serious Games Evaluation. As humans display their emotions and cognitive states through different channels, also using non-verbal communication elements like facial expressions, eye movements, body movements, tone of voice, posture and physiological reactions, pure textual data is not enough to fully capture the experience. In that context, researchers explore approaches to allow computers to

[2] http://www.google.com/analytics (last accessed January 2018)
[3] http://piwik.org/(last accessed January 2018)
[4] http://secondlife.com (last accessed January 2018)
[5] http://opensimulator.org (last accessed January 2018)

detect and recognize changes in a person's affective state. This is the specialty of the field of Affective Computing which is largely applied in evaluating digital products as one application possibility [206].

**Multimodal Learning Analytics**

The field of *Multimodal Learning Analytics* (MMLA) can provide useful techniques for aggregating and interpreting data captured along several modes. This new research area "emphasizes the analysis of natural rich modalities of communication during situated interpersonal and computer-mediated learning activities" [197]. For instance, measuring attention, effort and excitement within learning environments can benefit from biometric data like information about posture, facial expressions, eye tracking, pupil diameter, skin conductance, heart rate, respiration and EEG signals [133]. The author of [32] demonstrated how MMLA can be used for naturalistic assessments and real-time evaluation by analyzing programming snapshots, word counting, behavioral traces and video recordings of students to categorize them according to their programming style and expertise as well as their affect towards the learning subject, even without considering their test performance. A similar study [227] explored the combination of audio, video and digital pen writing signals for the identification of leaders and experts in a group. An approach for synchronizing eye-tracking data and log events with EEG signals was presented in [133], where they were imported as hits at significant timestamps of the EEG traces to help find correlations across channels. There are many challenges associated with this information-intensive approach. Not only can these types of data usually only be collected from a small group of people, but combining parallel streams of wireless data also produces complex and noisy results which are difficult to reduce and synchronize for analysis. On the other hand, this approach helps to identify significant events which might be overlooked using only extensive data. In addition, capturing rich modalities of communication using multimodal recordings (e.g. speech, digital pen, images and videos) can also be especially useful for evaluating collaborative learning settings [197]. In fact, traditional analytics based on clicks and keystrokes limit their integration into new technologies which no longer depend on mouse and keyboard for their input. An example of successfully combining multimodal data streams is described in detail in [32].

## 2.5 Mobile Sensing

Reports of the Serious Games market point to the mobile sector as the dominant future direction, especially location-based learning games and mobile augmented reality games as increasingly used paradigms [1]. Along with new types of user-generated data, the use of smartphones offers new contexts in which learning occurs which should be taken into consideration when evaluating users' interactions and also have direct effects and pose new challenges on other evaluation aspects such as affect detection [36]. This shift makes testing in traditional laboratory environments not only difficult but also not naturalistic as it ignores a large scope of the experience. As Learning Analytics based purely on logging of traditional interaction patterns becomes insufficient, it needs to be enriched with multimodal mobile sensing data in order to detect, for instance, that a learner is being distracted by environmental or social factors.

The term Mobile Learning Analytics describes "the collection, analysis and reporting of the data of mobile learners, which can be collected from the mobile interactions between learners, mobile devices and available learning materials" [3]. Thus, the field of Mobile Sensing, which is concerned with using mobile sensors to unobtrusively track users and their non-static environment, also started to play a growing role in the evaluation of Serious Games. The following are some smartphone sensors and their potential affective and cognitive measures listed in [235]: Smartphone Cameras can be used for eye-tracking and facial feature recognition which can be useful for inferring information about attention, stress, learning patterns, different skills and emotions. Accelerometer, gyroscope and compass can be used to detect body motion which could be used to measure arousal. Finger pressure on touch screens can be indicators of stress, certainty of response and cognitive load. Voice recorded by the microphone can be analyzed to measure stress and other affective states.

### Mobile and Ubiquitous Learning Analytics

Naturally, as soon as mobility and context come into play, learning analytics based only on pure logging of traditional interaction patterns becomes insufficient. In addition, the new modalities of interaction enabled by modern devices demand new modalities of tracing. An example of novel interaction paradigms are gaze-interaction games and apps where objects on the screen are moved using eye movements [30]. In fact, it would not be wise to ignore the rich multimodal data capturing capabilities of advanced technologies when they can offer deeper insight into learning interactions which can enrich the interpretation of log files for evaluation. Overlooking significant events, like a learner being distracted by environmental or social factors, can lead to drawing incorrect conclusions in evaluation studies [177].

Smartphones present a promising platform for learning analytics because of their widespread availability and ease of use, flexibility, multimodality and personalization which makes gathering data much more natural and non-invasive than traditional platforms. As the terms Mobile and Ubiquitous Learning are becoming more and more popular, new terms like *Mobile and Ubiquitous Learning Analytics* are beginning to arise [3].

Mobile Learning Analytics is concerned with data of mobile learners resulting from their interactions with mobile devices, learning materials and other mobile learners while Ubiquitous Learning Analytics additionally consider contextual information about learners and their environment such as time, location, activity, noise, light and social environment.

An example of an application which makes use of context information for learning is SCROLL (System for Capturing and Reminding of Ubiquitous Learning Log) [190]. Once the learner arrives at a place where s/he had a learning session before, s/he is reminded of the topic s/he learned at the same place. In addition, the learner can also select to view his/her own or other learners' learning log history of a selected time period or around a certain position. Another interesting question arises for learning systems which can be used on both PCs and handheld devices. The two versions of the software may not be identical, but rather complement each other, with one of them being recommended to the learner according to the current location or context [39].

The use of mobile technologies for learning on the move, referred to as mobile learning or m-learning, is becoming a well-established research field shaping the future of technology-enhanced learning [106]. Mobile devices can offer learning opportunities anywhere and anytime: Even in leisure-time people can benefit from entertaining mobile games for learning or improving different skills [111, 244]. As these technologies create new opportunities for learning, they also pose their own challenges, especially when it comes to evaluating effectiveness of different mobile learning practices [270]. Migrating Learning Analytics (LA) [17], from traditional settings, where they have proven successful [147, 251, 60], to mobile environments to make assessments in natural, non-stationary settings requires considering many new factors influencing the learning process like dynamic context, device capabilities and social interactions [268].

As mobile users nowadays use their devices throughout their day to accomplish different tasks and access a variety of services, many studies have been carried out to investigate the use of mobile devices to collect data about users and track their daily activities [43]. Following this trend, the term *Mobile Learning Analytics* was defined as describing "the collection, analysis and reporting of the data of mobile learners, which can be collected from the mobile interactions between learners, mobile devices and available learning materials" [4].

## 2.6 Conclusion

This chapter described some relevant terms and research fields associated with the research topic. In addition, the differnt steps involved in the evaluation process of serious games and how analytics can be integrated were outlined. For modelling knowledge and skills within Serious Games, suitable approaches like Narrative Game-Based Learning Objects (NGLOB) and Evidence-Centered Design (ECD) have been investigated. Choosing which data is to be logged in Serious Games depends on learning goals, set-

ting and tasks as well as game genre, mechanic and platform. Here it can be differentiated between intensive and extensive data, single-player and multiplayer games and generic and game-specific traces. Ideally, games should be initially designed in a way where all game mechanics reflect learner states to make learning analytics more efficient. As games are played on different platforms and with different interaction mechanisms, relying solely on activity logs will not be suitable for all kinds of serious games. This is why terms like *Multimodal Learning Analytics* and *Mobile and Ubiquitous Learning Analytics* have arised which have also been presented in this paper as well as different tools useful for aggregating data across users and modalities. Deriving measures from gathered data can be done by defining conditions on generic or game-specific variables and events which reflect aspects like learning, strategies and motivation. In order to be used in games, this analysis is required to be carried out and its results deployed and/or visualized in real-time.

Although there is an increasing amount of literature on this topic, there is a need for comprehensive guidelines, especially in the critical domains of interoperability, multimodality and mobility. Due to the complex nature of Serious Games, it was found that there is a need for defining theoretical frameworks especially tailoring research on Learning Analytics to this rich learning environment. As a first step, the overall process was described in Figure 2.3, also published in [240].

# 3 Related Work

Making sure that a Serious Game reaches the goal which was set by the teacher/developer is not an easy task. Not only should engagement and usability of an educational game be tested to make sure people will use it long enough without getting bored or frustrated, the learning outcome should also be assessed. Even with careful design and heuristic evaluation some issues are only discovered in user studies incorporating realistic settings. However, to justify the expenses of using a Serious Game in a learning context, the game has to be evaluated and its benefit has to be proven. This is why developing good frameworks for the evaluation of Serious Games is a topic addressed by many researchers [83].

A main focus of this chapter is to investigate the ways in which the field of Serious Games evaluation can benefit from multimodal methods by reviewing current trends as well as future potential. Although there are numerous reviews on Serious Games evaluation frameworks and techniques, they lack the focus on multimodality despite its many promises and challenges.

In this chapter, an overview over related research is given. First, studies using multiple modalities in evaluation sessions are reviewed to demonstrate the potential of using more than one modality in revealing valuable insight into gameplay interactions. Then, multimodal capabilities of smartphones are investigated with regard to their possible uses in Serious Games evaluation. Next, theoretical frameworks addressing the evaluation of Serious Games are compared as a foundation for the current research. Finally, evaluation tools and platforms suitable for aiding this kind of studies are reviewed.

## 3.1 The Use of Multimodal Data for Serious Games Analytics

Tables 3.1, 3.2, 3.3 and 3.4 present summaries of studies using multimodal data for the evaluation of Serious Games, learning environments, non-educational games and multimedia, respectively.

Indeed, most studies evaluating Serious Games, non-gaming learning environments and non-educational games using multimodal methods as seen in the tables focus on determining aspects from the fourth category: especially logging user actions and recognizing cognitive and affective states through different modalities, then combining both to examine aspects in the product assessed. By examining studies determining cognitive states, the order of modalities used according to the frequency of use in the listed studies is the following: Eyetracking was found to be the most used modality for determining cognitve states, followed by physiological data and lastly audio. Other modalities like pressure exerted on mouse and seat also played a role in one study in determining cognitive states. As for automatically recognizing affective states for evaluation, it was found that facial expressions is the most frequently used feature in the listed studies, followed by physiological data, then posture, gesture and finally audio. Most studies focusing on user performance and strategies included logging with the other multimodal data in their evaluation. Above all, the tables show how multimodal data are playing an increasing role in the evaluation of learning and gaming aspects and, while still limited, in the evaluation of Serious Games.

**Table 3.1:** A Summary of Serious Games Studies Using Multimodal Data.

| Study | Goal | | | | Aspects | | | Modalities/Features | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Evaluation | Adaptation | Comparison | Relation | Learning | Gaming | Using | Logging | Facial | Eyetracking | Posture | Gesture | Audio | Physiol. | Qualitative |
| [156] | X | | | | | | ineffective presentation and navigation | X | | X | | | | | questionnaires, think-aloud |
| [219] | | | X | | on- and off-task behavior and conversation, gaming the system | affective states | | | | | | | X | | live annotation |
| [180] | X | | | | | | attention to hints | | | X | | | | | |
| [55] | | | | X | | emotions | | | | | | | | HRV, EMG | GT: pre-post tests, logging |
| [200] | | X | | | attention level | valence, arousal | | X | X | | | | | the MindWave BCI, EEG | |
| [84] | | | | X | visual processing, problem-solving, and attention | | | X | | X | | | | | |
| [122] | X | | | | attention-grabbing elements | | | X | | X | | | | | |
| [42] | | | | X | | performance assessment | | | | X | | | | | GT: questionnaire |
| [136] | | | X | X | | pleasure | | | | | | | | HRV | |
| [14] | | X | | | | emotions | | | X | | | | | | |
| [191] | | | | X | fixation transitions and learning outcomes | | | | | X | | | | | |
| [56] | X | | | X | attention, motivation, vigilance | | | | | | | | | EEG | |
| [198] | | X | | | | affect | | X | | | X | | | | GT: live Observation using BROMP |
| [35] | | | X | | boredom, confusion | engagement, frustration | | X | X | | | | | | GT: live Observation using BROMP |
| [185] | X | | | | | pleasentness of game elements | | | | | | | | Startle EyeBlink (EMG) | |
| [110] | | | X | | off-task behavior | six affects | | X | X | | | | | | live annotation |

1

[1] Abbreviations: GT: Ground Truth, EEG: Electroencephalogram, EMG: Electromyography, HRV: Heart Rate Variability, BCI: Brain-Computer Interface, BROMP: a protocol for Quantitative Field Observations of Student Affect and Behavior

**Table 3.2:** A Summary of Non-Gaming Learning Environments' Studies Using Multimodal Data.

| Study | Goal | | | | Aspects | | | Modalities/Features | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Evaluation | Adaptation | Comparison | Relation | Learning | Gaming | Using | Logging | Facial | Eyetracking | Posture | Gesture | Audio | Physiol. | Qualitat. | Other |
| [277] | | X | | | performance assessment (boredom,confusion) | engagement, frustration | | X | X | | X | | | | GT: annotation by learner, peer and judges | |
| [20] | | X | | | | emotions | | | X | | X | X | X | | GT: annotated dataset | |
| [87] | | | | X | | embodied affect in dialogue acts | | X | | | X | X | | SC | GT: dialogue act tags | |
| [224] | | X | | | | valence | | X | X | | | | | HR, breath frequency, GSR, temperature | GT: expert annotation | |
| [2] | | | X | | boredom, confusion | delight, engagement, frustration, disappointment | | | X | | | | | | GT: annotation | |
| [15] | | X | | | | seven emotions | | | X | | | | X | | | |
| [88] | | X | | | learning | engagement, frustration | | X | X | | X | X | X | | GT: retrospective self-reports, pre-posttest | |
| [224] | | X | | | | | | X | X | | | | | HR, GSR, skin temperature, breath rate | | screen recording |
| [11] | | | | X | confidence, interest | frustration, excitement | | | X | | X | | | SC | | pressure on mouse, pressure on seat |
| [225] | | | | X | carelessness | affect | | X | | | | | | | annotation | |
| [37] | | X | | | total time, responce accuracy, proper function exection (on-task vs off-task) | engagement | | X | | X | X | | | | | |
| [72] | X | | | X | classifying dialogue acts | | | X | | | X | X | X | | GT: annotation | |
| [16] | | X | | | | emotions | | | | | | | X | | | |
| [45] | | X | | | attention and workload | | | | | | | | | EEG (The Emotiv device) | GT: debriefing | |

2

[2] Abbreviations: GT: Ground Truth, EEG: Electroencephalogram, GSR: Galvanic Skin Response, SC: Skin Conductance, HR: Heart Rate

**Table 3.3:** A Summary of Non-Educational Games' Studies Using Multimodal Data.

| Study | Goal | | | | Aspects | | | Modalities/Features | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Evaluation | Adaptation | Comparison | Relation | Learning | Gaming | Using | Logging | Facial | Eyetracking | Posture | Gesture | Audio | Physiol. | Qualitat. | Other |
| [216] | X | | | | | valence and arousal | | | | | | | | fEMG, SC, HR | | screen recording |
| [150] | X | | | | | boredom, challenge, excitement, frustration, fun | | | | | | | | fEMG, GSR, HR, EKG | GT: self-reports | |
| [274] | | X | | | | player experience | | | | | X | | | | | pressure patterns |
| [62] | | | | X | | player experience | | | | | | | | HR, EDA | | |
| [183] | X | | | | | boredom, immersion, and flow | | | | X | | | | EEG | GT: game experience questionnaire | |
| [232] | | X | | | | engagement, frustration, challenge | | | | | | X | | | | |
| [188] | | | | X | competence | flow, tension, challenge, positive and negative affect, immersion | | X | | | | | | HRV | self-Reports | |
| [41] | | | | X | attention | challenge, frustration, fun, anxiety, engagement | | | | | | | X | | GT: questionnaire | |
| [33] | | X | | | | neutrality, happiness, anger | | | X | | | | | | | |
| [186] | X | X | | | | valence and arousal | | X | | | | | | SC, HR, fEMG | | |
| [231] | | X | | | | happiness, sadness, anger, surprise | | | X | | X | X | | | | |
| [152] | | | | X | | play performance | | | | | | | | | | key graphical and sound gameplay feedback |

[3]

[3] Abbreviations: GT: Ground Truth, EEG: Electroencephalogram, fEMG: facial Electromyography, HR(V): Heart Rate (Variability), EDA: Electrodermal Activity , EKG: Electrocardiogram

**Table 3.4:** A Summary of Studies Using Multimodal Data for the Evaluation of Multimedia.

| Study | Goal | | | | Aspects | | | Modalities/Features | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Evaluation | Adaptation | Comparison | Relation | Learning | Gaming | Using | Logging | Facial | Eyetracking | Posture | Gesture | Audio | Physiol. | Qualitat. | Other |
| [69] | X | | | | mental states (agreeing, concentrating, disagreeing, interested, thinking and unsure) | | | | X | | | X | | | GT: labels on acted videos | |
| [23] | X | | | X | | emotions and engagement | | X | | | | | | HR, breathrate | | |
| [71] | | X | | | | 15 emotional states | | | | | | | | | GT: self reports | keystroke dynamics |
| [279] | X | | | | | user-perceived quality | | X | | X | | | X | | GT: Ratings, observation | |
| [21] | X | | | | | enjoyment, emotions | | | X | | | | X | | GT: final star rating | blink, shaking |
| [267] | | | X | | | valence and arousal | | | X | | | | | | Labelling, generic metadata | |
| [160] | X | | | | | “liking” and “desire to view again” | | | X | | | | | | GT: questionnaire | |

[4]

[4] Abbreviations: GT: Ground Truth, HR: Heart Rate

## 3.2 Challenges of Multimodal Evaluation

The process of extracting information from multimodal data presents many challenges. *Multimodal synchronization* is needed to prevent misalignments between different modalities due to different time scales [258]. *Multimodal Fusion* refers to the integration of multiple media (*data-level fusion*), their associated features (*feature-level fusion*), or the intermediate decisions (*decision-level fusion*) during the analysis of multimodal information.

A study on multimodal fusion in the field of multimodal learning analytics [282] compares three different approaches for fusing multimodal data for learning analytics and discusses them: Here, *Naive Fusion* is the approach used for conducting exploratory research by integrating aggregate features from different modalities without specific apriori assumptions about any interactions between them and ignoring temporal dependencies. The term *Low-Level Fusion* was used to refer to the approach where features are merged on very small time scales to uncover contextual dependencies during immediate actions which is also generally referred to as data-level fusion. In *High-Level Fusion*, the features are joined after classification, for instance to further examine different types of the same feature depending on another modality which adds a semantic level to the analysis but, on the other hand, may mask low-level differences in the data.

While most studies of multimodal fusion perform integration on fixed segments across all modalities, [138] investigates *Event-Driven Fusion* recognizing that not all relevant cues happen at the same time in all modalities. In this approach which is especially useful for real-time applications, information from each modality is used seperately to decide when more information is needed in other modalities, taking into account time shifts which can occur. For example, an emotional expression may first present itsself in voice before its facial reaction occurs. Thus, an event-driven approach may improve results of emotion recognition across different modalities.

Another problem with multimodal recognition which is being investigated in literature is the generalizability of recognition techniques to new populations and domains [221]. As Rowe's study indicates, not all features and techniques used for affect detection will yield the same results in different settings; in this case posture-based affect detection in learning environments was tested. Another study [89] showed differences in affective models across different student age groups for the same task and setting. [54] discusses a challenge which is faced in the analytics field in general which is interoperability. For our focus, which is the evaluation of Serious Games, the topic of interoperability plays an especially important role because of the heterogenity of educational games. In fact, even when serious games are comparable, differences in the evaluation process (like different activities, different assessment measures and different statistical analysis methods) can lead to incomparable results [5]. In addition, there is a need for a large dataset of serious games covering different topics and educational levels and deployed in different contexts to make evaluation techniques and results comparable. To address the heterogenity of data, analytics can be designed to extract generic information which can be applied to different kinds of games, but this might lead to losing depth while game-specific data are tailored to specific games and thus cannot be generalized.

Another challenge arises when trying to apply evaluation techniques to serious games played by a large amount of players over a longer period of time. [96] presents some approaches to address this and increase scalability of replay logs, for instance recording the experience of only a subset of the overall player population or using intensive multimodal methods only in the prototyping phase with a limited number of users, which is one scenario where multimodal evaluation methods are most widely used.

Because of these reasons, a hybrid approach is best in our view, where a Serious Games expert reviews multimodal information with the help of a tool which facilitates navigation, applies some pre-processing on the data to make it accessible for the researcher, but does not draw conclusions from the data by itsself.

## 3.3 Multimodal Sensing on Smartphones

This section investigates opportunities for the application of smartphone sensors to support Analytics practices. An overview of the existing studies on utilizing smartphone sensing for different purposes which can equally be used for serious games analytics is given. In addition, the challenges associated with utilizing and analyzing multimodal smartphone sensor readings which should be considered when researching in this field are explored. Surprisingly, despite the evolving capabilities of smartphones and the maturity of mobile learning research, few studies have been carried out demonstrating the use of multimodal data using smartphones for Learning and/or Gaming Analytics. However, different studies showed the use of smartphones for eyetracking, facial feature extraction, voice analysis and other techniques useful in recognizing congition states which are considered valuable for Learning and Gaming Analytics.

### 3.3.1 Smartphone Sensors and Analytics

Smartphones nowadays are equipped with an increasing number of sensors such as GPS, accelerometers, gyroscopes, microphones, cameras, proximity, light, temperature and humidity sensors as well as radio antennas. In addition, other interactions and services on smartphones like sms, typing and swiping speed, touch type, strength and count and different key pressing frequences can also offer rich information for analytics [215].

Contrary to body-worn sensors, embedded smartphone sensors can collect rich information without being obtrusive and allowing natural interaction without requiring additional hardware. *Sensor-less Sensing* describes this "opportunisitc use of existing sensors embedded in daily use computing devices to repurpose their signals to track different biometric states representative of mental or physiological states directly or indirectly." [199]. Using smartphone sensors is a good example of this non-invasive sensing, due to them being widely used as integral to everyday life, and their connectivity enabling natural data collection. The author of [235] lists the following smartphone sensors and their potential physiological cognitive measures: Smartphone Cameras can be used for eyetracking and facial feature recognition which can be useful for infering information about attention, stress, learning patterns, different skills and emotions. Accelerometer, gyroscope and compass can be used to detect body motion which could be used to measure arousal. Finger pressure on touch screens can be indicators of stress, certainty of response and cognitive load. Voice recorded by the microphone can be analysed to measure stress and depression. A useful tool for real-time visualizing smartphone sensor readings to support the development of sensor-based applications is Sensorendipity [145]. In the following, it will be discussed in more detail how different smartphone sensors can be used for collecting information which can be useful for learning analytics and what challenges are associated with them.

The front-facing cameras of smartphones can be used for a variety of techniques to measure cognition which can also be used in learning analytics, like eyetracking and facial feature extraction, despite their generally lower resolution in comparison to the back-facing cameras [104].

Eyetracking techniques can give valuable insights for user evaluation techniques complementing traditional logging, e.g. how long a user watches a stimuli before making a selection or which hotspots particularly attract users and can thus be powerful in guiding educational design decisions. A calibration is performed for each user locating and recording the pupil while the user looks at different spots on the screen. This calibration information is used combined with real-time locations of pupils to determine the gaze location. In game development eyetracking can be used to measure player performance, engagement and visual attention leading to the introduction of new metrics and design principles as well as more efficient graphics algorithms optimizing perceived quality [260]. In [121], the results showed that eyetracking can provide important information about game based learning process and game designs. The authors of [259] propose a set of gaze visualization techniques for supporting gaze behavior analysis in virtual environments using advanced 3D scan paths and 3D attentional maps.

In [129], eyetracking using the smartphone camera was used to record and analyze the number of words and the types of documents users read, where they read them, how fast they read as well as their periods of concentrated reading, revealing quantified information about their reading habits. Additionally, users can compare their tracked reading habits with their friends. In this app, when the user scrolls to the end of an article slowly, at a threshold around 700 words per minute, while looking at the screen, the article is counted as read and the reading speed is recorded. Using Latent Semantic Mapping with a number of default categories which can be adjusted or extended by the user, the genre of the article is determined. To make the eyetracking and gaze estimation techniques less dependent on the calibration phase, lighting conditions and orientation of the device, the authors suggest using light and motion sensors and take this obtained information into consideration.

The authors of [48] used a smartphone mounted on the dashboard of a vehicle with front-camera facing the driver to estimate driver's gaze direction during driving. Gaze classification was carried out with respect to eight gaze directions specified apriori for efficient implementation. Wide variations across different drivers, vehicles, smartphones and camera placements have been identified making this classification method not well generalizable, which is why the authors propose initially gathering training data for each driver/vehicle/camera setup combination. As using controlled gazing during driving is unsafe, the proposed technique consists of first carrying out training while the vehicle is stationary and then combining this with data from the moving vehicle with simple road gazing to account for camera jitter, driver movement, varying illumination, variability in driver pose etc.. Using a variant of this method with only two broad gaze categories, safe (road, left mirror, right mirror, top mirror) and unsafe (dashboard, phone/text, music console), very high accuracy has been obtained which can be used for monitoring safe and unsafe driver behavior. Similarly, mobile pilot behavior was monitored using eyetracking in [278].

Facial images have been used with good accuracy for emotion detection [146], differentiating six main categories of facial expressions (surprise, fear, sadness, angry, disgust, and happiness) [68]. Extracting facial expressions using the smartphone's front-facing camera and correlating them with context, like checking an sms, can help infer emotion and affect in everyday scenarios. For mobile learning, this can be used to infer information about learner engagement. In fact, studies showed that smartphone cameras can be used with good accuracy for facial expression recognition, especially as the computing performance of smartphones is increasing [46, 151]. The authors of [46] propose a robust facial expression recognition system on a smartphone using Active Appearance Model (AAM) fitting methods and backpropagation neural networks. In [151], a framework to track user's emotional engagement to videos played on a smartphone is presented by correlating facial expressions to video content. For this, CSIRO Face Analysis SDK (CFAS), a library for facial geometry was used for tracking facial landmarks [57].

As most users hold their smartphones close to their lower half of the face while watching videos, attributes for lip corners were found to be more suitable for engagement monitoring than attributes for eyes and cheeks due to their bigger relative size on the captured videos. It was also observed that some postures of users while using smartphones doesn't provide their face in the field of view of the front-facing camera.

*Appearance Context Logging*, implemented in [285] is another use of the smartphone camera which automatically and opportunistically captures face images through the day. The idea is that spontaneous camptures track more naturally how the user is perceived by others. The FaceLog app takes a picture of the user's face when he unlocks the screen of his phone. As most of the pictures obtained using this method had bad quality due to motion, posture, camera viewing angle or low brightness, the authors suggest using predictive filtering before activating the camera by using other smartphone sensors like accelerometer and luminosity sensors to rule out moments resulting in such low quality pictures and thus save energy.

In addition, moments of unlocking the screen turned out to be less appropriate for capturing facial expressions as most pictures contained an expressionless face. The authors suggest choosing moments of specific user-to-phone interactions like checking messages.

Another use of facial images is the extraction of vital signs from face images such as pulse detection through color changes due to blood flow [210] and breathing measurement through small movements on the chest. Both approaches have been realized in Philips "Vital Signs Camera" app for the iPad. As such signs correlate to arousal, they can also be used in the evaluation of learner experiences.

All mobile phones have built-in microphones which can be used for voice analysis. Studies have shown modest accuracy at measuring emotion and high accuracy at estimating stress [44, 144]. The toolkit AMMON [44] has produced high accuracy emotion detection through the detection of voice changes, pauses and speech speed using smartphones. StressSense [144] was developed to detect real-time stress from real-life conversations using android phones and its robustness was proven across different individuals in different indoor and outdoor environments. This was tested during job interviews and outdoor job execution tasks and compared to indoor and outdoor reading recordings. Stress readings from GSR sensors were used as ground truth. Two adaptation schemes (supervised and unsupervised training) were used to efficiently adapt a universal stress model to different users and scenarios using few observations. Supervised adaptation was found to be a practical alternative to a personalized scheme, when limited training data is available for a new user. Using smartphone microphone data to enhance learning experiences may for instance especially be useful when children are using mobile learning apps as they tend to express their joy and confusion verbally.

Instead of mouse and keyboard, users of smartphones and tablets predominantly use touch interactions with touch strength and movement introducing new sources of sensory data. The authors of [233] demonstrate the ability to predict student effort level by logging characteristics of a student's touch interaction while solving mathematical problems on a tablet and correlating them to high and low effort problems. Students using the touchMath iPad app input their answers through writing/drawing their working steps in an empty space beneath each questions. Stroke points, movements and pressure are recorded as long as the student touches the screen providing information about stroke time, distance and velocity. Touch pressure is determined using the iPad's accelerometer detecting positive and negative movements along the z axis with a sufficient sensitivity comparable to pressure-sensitive tablet screens. Students are also asked to self-report their perceived correctness of their delivered solutions. Images of each solution and related touch data are uploaded to the server along with problem level information for later analysis. The study showed that touch pressure on a tablet screen can be predictive of the level of student effort. The authors argue that this can be important for adaptive learning environments to guide more accurate interventions which help avoid frustration. The authors also plan to add erasing interaction as a measurement of affective states.

A system for logging, analyzing and visualizing all touch operations on an Android phone was presented in [99]. After obtaining touch logs with timestamps and current applications, touch operations (single/multi touch, single/multi swipe, pinch in/out and rotate) were recognized, swipe speed was calculated and results were visualized. Touch logs recorded automatically on Android OS already contain timestamps which are relative to latest wakeup and has slightly different formats on different terminals and were thus converted into a common format: absolute unix time. The current application was recognized using a system dump. Touch behavior was found to differ according to person and context even using the same application. For the visualization, each touch was drawn by a circle whose radius depends on touch pressure and whose color corresponds to the application used.

A popular use of GPS and accelerometers is to track movement and physical activity using heallth-promoting apps [90]. Recording places we visit in lifelogging apps was also used to aid memory [112]. An example for using location information for learning is SCROLL (System for Capturing and Reminding of Ubiquitous Learning Log) [190]. Once the learner arrives at a place where s/he had a learning session before, s/he is reminded of the topic s/he learned at the same place. In addition, the learner can also select to view his/her own or other learners' learning log history of a selected time period or around a

certain position. The detection of common gestures using accelerometer and gyroscopic data in smartphones has been investigated in [199]. For instance, oscillation of the arm during common task such as dialing, searching and browsing can give clues which can be used for activity recognition. In [104], the authors propose connecting smartphones to external sensors in close proximity via bluetooth to acquire sensor data. For this to function, both mobile and sensor device must support the same protocol stack and have compatible bluetooth versions. The author of [114] demonstrates an unobtrusive mobile electrocardiogram (ECG) monitoring system that monitors the user's ECG opportunistically during daily smartphone use allowing inferences about stress, emotion, and even sleep quality using a prototype sensor device attached to the test smartphone.

### 3.3.2 App Use and Phone Interactions

In addition to acoustic data and facial expressions, affect can also be measured on smartphones using phone interactions and app usage [215]. The number of sms sent, the number of calls received and made, the duration of calls, the number of presses on backspace/enter key on the screen as well as social interaction patterns using Bluetooth and Wi-Fi scan and GPS traces can be used as usage clues. Accuracy of affect sensing using this kind of data is lower compared to audio and video captures due to lower correlation but are much less invasive [215]. However, they can provide good results when used for app recommendation [235]. For instance, the study in [34] investigated the correlation between app usage and time and concluded that time is highly related to app usage behavior. A time-based apps predictor was proposed in [137] to dynamically predict the apps which are most likely to be used by deriving probability models through extracting global, temporal and periodical usage features from app usage traces. Apps with highest usage probabilities at a certain time are recommended to the user. Apptracker [26] demonstrates logging app usage across multiple devices (iOS, Android and OS X). It logs the time when apps are opened and closed and when devices are locked and unlocked. The authors point out that an app being open does not mean it is in use, as some apps like facebook can remain open in the background, and that categories used to order app stores are not accurate in describing use as there are considerable differences between usage patterns in apps falling under the same category, such as different social networking apps.

### 3.3.3 Applications

This section will first examin different successful applications of user profiling using smartphone sensing will and then discuss possible implications for serious games analytics. The author of [26] points out that the patterned ways in which mobile devices are used through daily life have made it possible to infer information about people's lives such as when they are in bed or getting home from work by just observing phone data like when a device is in use. Combined with other information like device charging, Wifi information or apps like alarm, map, transportation timetables or taxi-booking services, more accurate pictures of life can be depicted. Lifelogging services which attempt to track our everyday experiences and activities like sleep, exercise, food, mood, location etc. with the aim of tracking progress and mediating memory are becoming increasingly popular [218]. Experiments in this direction are carried out by the *Quantified Self* community, where users and developers collaborate to explore self-tracking methods of making our lives measurable.

Several applications use real-time emotional and stress estimates using smartphones to offer *Adaptive Intervention*. In [199], interventions based on social and gaming concepts are applied to regulate stress. At times with significant changes, the user is also prompted to give a self-report which can be further used for training. Similarly, a smartphone-based social alert system for autistic children [47] records and analyzes data using smartphones carried by autistic children to automatically detect or predict their stereotypical behaviors and alert teachers or caregivers enabling early intervention. Predicting is implemented by correlating environmental factors with triggered behaviors using smartphone sensors like

accelerometers and microphones and has demonstrated high accuracy. Annotations made by caregivers on a remote server further aid classification.

Audience research has also benefited from mobile sensor readings [178] by tracking museum visitor experience to reveal changing dynamics visitor behavior over time. The authors recognize advantages of mobile sensing in this scenario as reduced effort of researchers, reduced time and cost of data collection, increased accuracy and performance of tracking visitors' positions as well as advanced analytics opportunities. Another application field of using information obtained from smartphone sensors is *Augmented Cognition* where cognition activities are modified according to physiological sensor feedback to increase task performance. The authors of [104] discuss issues involved in implementing augmented cognition activities on a mobile platform and the tradeoffs of gamifying augmented cognition activities. A ubiquitous Augmented Cognition framework is proposed in [251].

In Mobile Learning, context-aware, personalized and adapted applications can be developed which adapt learning activities according to factors like contextual information, learner attributes and real-time affect. The authors of [289] outline that a mobile delviery tool is required to be able to automatically detect and let the user input contextual information, to enable adaptation mechanisms on the delivered educational activities, resources, tools and services according to the context, to enable execution of learning activities without an internet connection after adaptation and to provide a visualization for easy navigation of learning activities. An example of an application which makes use of context information for learning is Units of Learning mobile Player (UoLmP) [85] where the mobile context of the learner is described by his/her temporal information including mood, interests, needs and preferences, other people related to current context who influence the learning process, current accurate location, cultural background and learning setting, technological and non-technological learning tools, current time conditions and constraints as well as physical conditions. This information is used by the system to adapt its learning flow. This was applied to an exemplary learning scenario in a language-learning center involving collaborative learning activities with real-life tasks. In [218] the authors propose using data offered by Quantified Self approaches to support the improvement of learner's reflective learning experiences.

Personalization of learning materials can also be achieved by adapting content to the user's device capabilities and UI preferences. A tool is described in [149] which automatically gerenates multimodal web applications for a target device by using intermediate model-based representations of the user interface's logical structure. The user interface does not only adapt to properties like zooming level chosen by the user but also responds to changes in environmental factors like noise and light conditions. Another interesting question arises for learning systems which can be used on both PCs and handheld devices. The two versions of the software may not be identical, but rather complement each other, with one of them being recommended to the learner according to the current location or context [39].

As affect detection in an intelligent tutoring environment has already been proved to improve learning effectiveness [10], we argue that collecting multimodal data from smartphones offers unprecedented opportunities for the design of adaptive learning games and applications on mobile devices. For example, learner engagement in mobile learning games can be used for adapting game elements or leaner attributes and preferences collected through lifelogging can give insights helping learning app and game genre recommendations. Measuring attention and effort during a mobile learning session using methods described in Section 3.3.1 can provide valuable data for learning analytics. Multimodal approaches can further enhance accuracy by fusing sensor readings to obtain meaningful results.

### 3.3.4 Overcoming Smartphone Limitations

Many algorithms of learning analytics require high processing, storage and energy capacity for data collection and analysis which are obviously limited on mobile devices, especially when increasing sensor readings and processing of multimodal data, [127]. However, not only are capabilities of mobile devices on the rise following Moore's law, but there are also other opportunities offered by smartphones which

can be exploited to cope with or even eliminate these challenges. Mobile cloud computing and storage allows computing and data storage to be off-loaded to the cloud through constant wireless connectivity of smartphones which is becoming mainstream. This also helps reduce power consumption.

In [253], a phone-cloud collaboration model is proposed where the phone acts as a sensory organ while the cloud acts like a brain. Sensor readings and activity detection results are sent from the phone to the cloud where they are used to train the phone-independent probabilistic models associated with the user and as fed into prediction algorithms. Data transmission load is low as only results are communicated. Additionally, for increasing battery life for always-sensing, a smartphone design is proposed where the core network is isolated from the sensor network to allow CPU and main bus to sleep while independent sensor subsystems are working. The sensor peripheral in this design has its own processing units enabling inter-sensor communication, sensor data manipulation and fusion, with the size of each subsystem depends on its requirements.

A crucial factor influencing computational speed and power consumption is algorithm efficiency, like using low-power sensor readings to rule out the need for computationally intensive algorithms and more power-intensive sensors where possible. For classification and reconstruction of sensor data, *Compressive Sensing* [22] is also a promising theory which can address resource constraints on smartphones.

Other challenges associated with capturing and extracting information from multimodal recordings on smartphones are related to the unconstrained mobile nature of these devices which requires the study of various factors of the dynamic physical environment. Facial feature extraction using smartphone's front-facing camera can suffer from unfavorabble conditions such as illumination and exposure variations, user distractions, unstable camera and low capturing quality [151]. To overcome these challenges, other smartphone sensors can provide information supporting adaptive recording Sensor-fusion can not only improve quality of photos and recordings, but can also support analysis, leading to more accurate inferences. In [46], the application of the Difference of Gaussian (DoG) kernel obtained good results under different illumination conditions improved results of facial feature extraction.

Similarly, voice analysis of phone recordings suffers from background noise, special attributes of sound during phone conversations and acoustic attributes being lost due to dut to automatic gain control and noise cancellation using the phone microphone. Here, readings from other sensors can also help in opportunistic recording and signal reconstruction. Privacy concerns further complicate collecting sufficient data for training voice analysis models [144]. This issue is discussed in the next section. As media file sizes on smartphones are smaller than PC video files, their analysis are less computationally intensive and thus naturally faster [151].

### 3.3.5 Privacy Concerns

A considerable challenge associated with gathering data about smartphone users is concerned with ethical considerations as collecting and disseminating sensor data raises serious privacy and security issues [163]. Sharing sensitive information might lead to information or even device theft [74].

To address this, the design of such systems should require data to be stored and aggregated in such a way that ensure users' privacy [47, 26]. For instance, performing computation on the cloud requires it to emphasize privacy management [253]. Users should also be able to opt out of logging services at any time.

In [265], authors investigated reducing the privacy infringing content of first-person point-of-view images taken by wearable cameras while still maintaining evidence of everyday eating behaviors. The balance between images that might pose a privacy concern versus images that contain information salient to a particular interest were quantified and made evident using a privacy-saliency matrix representation. Using this model, techniques like face detection, cropping, location filtering and motion filtering were applied to address privacy risks.

In addition, other sensor data was used to recognize activities of interest to discard photos likely to introduce privacy issues. On the other hand, mobile multimodal learning analytics can also be used in

controlled user studies with participants agreeing to share their information for the purpose of evaluation.

## 3.4 Theoretical Frameworks for Serious Games Evaluation

Evaluating Serious Games requires considering different dimensions. As described in [161], playful learning applications should be assessed in terms of their playability, learning and usability. As Serious Games, per definition, are required to be at least more engaging than regular learning environments, proving that a Serious Game is captivating and motivating, irrespective of its learning goals, is essential. However, a Serious Game which entertains but does not meet its educational goals is similar to a medicine which tastes good but lacks the active ingredient. Proving the educational effectiveness of a Serious Game can be complicated, depending on the competency it is designed to convey.

Usability and User Experience are crucial for any kind of application and in Serious Games they are even more important [175]. Usability in this context is "the degree to which a player is able to learn, control and understand a game" [207]. Usability issues can hinder interaction, making all development efforts and costs useless. When surveying available frameworks for Serious Games Evaluation, four main roles arise: The user, the Serious Game, interactions between both and the context in which the interaction takes place. Evaluation frameworks which consider only one of these roles can have one of the following aims: Considering only aspects of the serious game without interactions with the user results in heuristic frameworks defining quality metrics which can be evaluated by mere inspection without the need for real users. Focusing on user behavior or learning without correlating with product qualities is usually used for assessing the learner and not for assessing the Serious Game. For this reason, we will focus in this section on introducing frameworks which cover at least aspects of the user (also referred to by some authors with terms like the "Serious Player" or the "Play-Learner") AND aspects of the product, here the Serious Game. Table 3.5 gives an overview over these frameworks and the dimensions considered by each. Under each dimension, the aspects are listed which emerged from the literature review as necessary to be considered when evaluating Serious Games. These are:

- learning, gaming and interaction design on the product side
- their counterparts: learner-, player- and user model on the user side
- coherence in design between the three dimensions
- in-game interaction metrics: *performance* (actions related to each of the three dimensions learning, gaming and interaction), *experience* (affect and cognition during play as a result of each of the dimensions)- see next section for more information and *strategy* (reason behind the actions which can also be inferred as we will see later in the model)
- context

A cell with an x in brackets means that this aspect is only partially addressed by the presented framework.

**Table 3.5:** Theoretical Frameworks for Serious Games Evaluation.

| Framework | Citation | Serious Game | | | | Context | Play-Learner | | | In-Game Interactions | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Learning | Gaming | Using | Design Align-ment | | Learner Model | Player Model | User Model | Performance | Experience | Strategy |
| A game-based learning framework | [269, 76] | X | X | | | X | (X) | (X) | | X | X | |
| Games, motivation and learning | [79] | X | X | | | | X | X | | X | X | X |
| Playful learning: An integrated design framework | [208] | X | X | X | | X | | | | X | X | |
| Design, play and experience framework | [281] | X | X | X | X | | (X) | (X) | (X) | X | X | |
| Conceptual research model for SG evaluation | [158, 159] | (X) | (X) | | | (X) | X | X | X | X | | X |
| EFM Model for Educational Game Design | [254] | X | X | | | | | | | | X | |
| Flow framework for assessing the quality of educational games | [123] | X | X | X | | | | | | | X | |
| Three-layered model | [78] | X | X | X | X | X | X | | | X | X | |
| SG design assessment framework | [173] | X | X | X | X | | | X | X | | | |
| The education and enter-tainment (E/E) grid | [167] | X | X | | | | X | X | | | | |
| The six facets of Serious Games | [155] | X | | X | X | X | | | | X | | |
| RAGE evalution framework | [256] | X | X | X | | (X) | (X) | (X) | (X) | X | X | |

## 3.5 Evaluation Platforms

In this section we will review tools suitable for helping in the process of evaluating Serious Games. In addition to tools specifically created for Serious Games, we will also shed the light on evaluation tools which can be used for any software or learning environment with special focus on tools with multimodal support.

### 3.5.1 Serious Game Analytics Platforms

The two major tools which were built for supporting prototyping and evaluation of Serious Games in literature are StoryPlay [165, 217] and GLEANER [98]. StoryPlay is a tool for collecting and visualizing learning and gameplay traces based on the Narrative Game-Based Learning Object (NGLOB) model [81] with which the Serious Games authoring environment StoryTec was created. This model represents a player's competencies, his/her player type as well as the narrative stage reached in the Serious Game to help online monitoring and adaptation of the three aspects. Thus a serious game (or a "story") created with StoryTec can be tested in real-time or offline using StoryPlay, whereby the researcher, instructor and/or game designer can view a replay of the gaming session alongside visualizations of any updates in the model and variables as well as aggregate traces of multiple players.

GLEANER or "Game LEarning Analytics for Educational Research" is described as a game and learning analytics tool to support serious games research [98]. It aids in choosing the proper data to be captured as well as in data analysis and communicates with game engines via http. Data collected in GLEANER include generic traces defined in [230] like game traces, phase traces, meaningful variable traces and input traces as well as aggregations of these traces. It also allows the construction of game-level and genre-level traces as aggregations of low-level traces. Another logging framework for educational games built in the Unity environment is referred to in [95] as Unity Logger. Unfortunately, we have not found any Serious Games Evaluation Tool which helps in synchronizing and/or analysing multimodal information.

### 3.5.2 Game Analytics Platforms

Playtesting is an integral part of game development which aids in evaluating and improving video games in different development stages. However, especially in the analysis of playtesting data, there is no consensus on the most effective methods. Most Game Analytics tools rely on logging or recording so-called game telemetry and then making aggregations and visualizations on the raw data to make it accessible for developers.

As described in [70] and [275], using Game Analytics can help developers improve game design and balance, "detect hidden problems and bottlenecks, categorize game contents and types of players as well as identify possible future monetization opportunities." Microsoft's TRUE system [124] collects gameplay, video and self-report data and visualizes metrics to help developers improve game mechanics. One research-based tool for aiding Game Analytics is Data Cracker [164], a "visual game analytic tool" to help monitor online multiplayer game-play before and after game release. HeapCraft [179] is a suite of tools for the game Minecraft helping to measure collaboration among its players with the aim of fostering strong player communities. For exploring arbitrary games, Gamalyzer [195] is a tool for clustering and visualizing sequences of player traces.

Evaluation sessions, especially in the game industry, often include mixed methods using qualitative and quantitative measures to help developers get a complete picture of player experience [263]. This is a challenging and time-consuming task which reveals the need for tools to help developers to efficiently combine multimodal data during and after evaluations [287, 263]. In [187] a tool is presented which automatically identifies game-related physiological reactions by logging in-game events, annotating physiological responses and synchronizing their readings with gameplay session videos. A similar

tool was discussed in [157] which combines logging, video, physiological data and self-reports. This approach of unifying quantitative and qualitative data has been investigated in [171], where the player's path in the game was visualized along with physiological measurements.

Aside from physiological measures, a tool for recording eye tracking data and relating it with gameplay logging, LAIF, was presented in [184]. A tool called Vixen [63] which came to our attention after our initial comparison of tools in [238] is used for interaction visualization of gameplay experiences. This tool might be the closest approach to our platform but is intended for general games created with unity and not especially for Serious Games as there is no learning analytics element to it. In addition, it does not run on mobile devices.

### 3.5.3 Interaction Analysis Platforms

Fusing multiple streams of sensor data is a complex task requiring synchronization and aligned representation. Chronoviz [75] is a tool which was introduced specifically for this purpose offering integrated and interactive visualization of collected multimodal data streams to support data exploration as well as in-depth analysis of significant moments. It provides several tools for interactively navigating, segmenting and annotating heterogeneous data from diverse sources such as multiple video and audio files, computer logs, sensor readings, paper notes, and transcriptions. Support for mobile eyetracking into this tool has also been studied in [278].

A similar tool is Replayer [177], a cross-platform and cross-language toolkit offering combined visual exploration of diverse recorded data such as media clips and statistical log data using an extensible distributed network system, where all inter-component communication takes place over TCP network connections. In [266], it is argued that multiple streams of data need to be structured to facilitate the analysis of complex collaborative learning environments by enabling a better understanding of effects of social interactions, task design and learning tool use on learning outcome.

In the field of user experience evaluation, there are tools which support data collection and analysis of user interactions with software in general, which can also be used for the evaluation of learning, gaming and other domain-specific applications. Here we will focus on tools supporting multimodality. In literature, these tools are referred to using different terms like Qualitative Data Analysis Software (QDAS), remote usability evaluation tools, session recording tools or interaction analysis tools.

In Tables 3.6 and 3.7 we considered only interaction analysis tools which support recording AND playback of multimodal data and which support logging or screen recording of interaction. This is why tools created for the evaluation of Serious Games which were discussed in Section 3.5.1 like the original form of the research tool StoryPlay were left out of this table as these tools do not support the recording and playback of multimodal data. Most Multimodal Tools for Game Evaluation which were presented in Section 3.5.2 like LAIF support either only recording or only playback. Also, most interaction analysis tools were left out as they do not support logging or screen recording of interaction. The same tools are used in the first column of all three tables but the originally one table was split in three tables for better readability of the columns. The different categories in the tables emerged as a result of the survey. In Table 3.6, the columns "Dimensions" and "Features" are presented: Here we differentiate the three dimensions Learning, Gaming and Interaction according to the focus and application field of the tool [5].

In Table 3.7, the tools are compared according to the modalities they support. As shown in the tables presented, most of the tools are pure interaction analysis tools with no focus on learning or gaming: Only two tools listed are created for game evaluation while only one is for learning. This confirms our claim that there is no tool especially created for the evaluation of Serious Games which considers the three dimensions of learning, gaming and interaction as well as enables recording and playback of multimodal data in addition to logging and replay. This re-search gap motivated our research to supplement the available literature.

[5] all web-links in the table were last accessed January 2018

**Table 3.6:** Evaluation Focus and Features of the Different Platforms.

| Tool | Citation | Dimensions | | | Features | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Learning | Gaming | Using | Synchro-nization | Navigation /Query | Pre-processing | Annotation | Analytics |
| ColAT | [12] | X | | X | X | X | X | X | X |
| Annotation Tool | [187] | | X | | X | | X | X | X |
| FBGM | [153] | | X | | X | X | X | | X |
| DRS | [38] | | | X | X | X | | X | X |
| Noldus Observer XT | [290] | | | X | X | X | X | X | X |
| Ogama | [273] | | | X | X | | X | X | |
| Mangold INTERACT | mangold-international.com | | | X | X | X | X | X | X |
| Silverback Guerilla | silverbackguerilla.com | | | X | X | | | X | |
| UXCam | uxcam.com | | | X | | | | X | X |
| Lookback | lookback.io | | | X | | | | X | |
| Playtestcloud | playtestcloud.com | | X | X | | X | | X | |

### 3.5.4 Playtesting Platforms for Mobile Games

As for mobile games, Playtestcloud [6] is an online playtesting platform which offers a software which wraps around a mobile game to equip it with screen and touch recording features without the need to modify the game or to integrate an SDK. They offer acces to playtesters, who, before they start the game, will see screens with instructions that walk them through tasks they have to accomplish for the playtest and prevent them from launching the game after the playtest has concluded. The software will record the screen contents of the app, all touch gestures and the microphone input.

In [246], the mentioned different research-based game, learning and interaction analysis tools were compared based on their features and it was noted that there was no such tool tailored specifically for Serious Games. This analysis underlines the need for including all important features found in these platforms into one Serious Games Analytics Platform. Next the different features will be described and how different publications of other tools have explained their importance.

For example, a tool called Vixen is described in [63] which enables interaction visualization of gameplay experiences. This tool is for general games created with unity and not especially for Serious Games as there is no learning analytics element to it. In addition it does not run on mobile devices. Like in other interaction analysis tools, the feature of recording and replaying the face expressions of game players is presented as having many advantages for evaluation: It is an unobtrusive way of observing players' engagement and involuntary reactions during playtesting [61]. Recognition and analysis of these expressions can also be applied on the recordings as a quantitative assessment method [262].

[6] http://www.playtestcloud.com (last accessed January 2018)

**Table 3.7:** Modalities Supported by the Different Platforms.

| **Tool** | **Modalities** | | | | | | |
|---|---|---|---|---|---|---|---|
| | Logging | Screen Recording | Video | Audio | Eyetracking | Physiological | Mobile |
| ColAT | X | | X | X | | | |
| Annotation Tool | X | X | | | | X | |
| FBGM | | X | X | | | X | |
| DRS | X | | X | X | | | X |
| Noldus Observer XT | X | X | X | X | X | X | X |
| Ogama | X | | X | X | X | | |
| Mangold INTERACT | X | X | X | X | | X | X |
| Silverback Guerilla | X | X | X | X | | | |
| UXCam | X | X | X | X | | | X |
| Lookback | | X | X | X | | | X |
| Playtestclowd | | X | | X | | | X |

Another publication about an annotation tool [187] describes that the use of a session video view with adjustable speed is desirable for investigating in-game action in parallel with their responses. Also allowing the user to quickly skip through a video to jump to a certain event or a certain reaction by clicking on this particular interest point was found to be a very useful feature. In StoryPlay, this is achieved not by recording screen video which would be big in size along with the front-facing camera video and not easy to navigate to a certain event. It is achieved by a session replay tool which reconstructs gameplay from the log files using the game engine itsself, as we will discuss later.

As for mobile games, Playtestcloud (playtestcloud.com) is an online playtesting platform which offers a software which wraps around a mobile game to equip it with screen and touch recording features without the need to modify the game or to integrate an SDK. They offer acces to playtesters, who, before they start the game, will see a screens with instructions that walk them through tasks they have to accomplish for the playtest and prevent them from launching the game after the playtest has concluded. The software will record the screen contents of the app, all touch gestures and the microphone input. StoryPlay has a similar approach in recording log files on mobile devices, as well as front-facing camera video and/or microphone recordings.

The feature of filtering out uninteresting video frames using information from low power mobile sensors like the illumination sensor was described in [94]. They also investigate predicting whether a frame contains faces using the accelerometer and gyroscope sensors. These features have been included in StoryPlay as well and this was one of the reasons why mobile sensor data is used in our framework. The other reason is that they also provide data on how the player is holding the device which is also important for UX testing.

In addition, location sensors have been used in some multimodal interaction analysis tools supporting mobile deployment to track the location of the testers during a remote testing session [77]. This is useful when having playtesters from different places and you need to gather location data and was thus also included in StoryPlay Multimodal.

The synchronization and interactive navigation of multimodal data is applied in many multimodal data analysis tools like ChronoViz [75], Tatiana [66], Digital Replay System [38], Noldus Observer [290] and Mangold Interact (mangold-international.com). It enables researchers to jump to the point of interest

and see all related multimodal data run simultaneously next to each other which saves time and effort of analyzing and annotating qualitative data.

## 3.6 Conclusion

This chapter first reviewed related work on modalities used in Serious Games evlauation studies. From the presented summarized studies, it is difficult to make conclusive statements about which multimodal data are more useful for evaluating which serious game elements. Every study only picked a method and an element without having a connection to other studies and cannot be generalized to other games. Another aspect which needs further research is how to facilitate closing the feedback loop and help developers infer recommended game-based learning design improvements from evaluation results. Although from the shown studies, multimodal data was captured mainly for measuring affective and cognitive states and thus cannot be directly used as a measure for learning outcomes, these states accompanying the learning process have been shown in learning theories to have a considerable effect on learning. In fact, the aim of using multimodal data is not to replace classical assessment of learning or even logging-based mechanisms of assessment but rather to enrich these methods where more data is needed and beneficial. Next, the literature review covered theoretical Serious Games evaluation frameworks and available tools, also published in [246, 238], and research uses of Smarthpone sensors and their challenges. Furthermore, a comparative analysis of features of available research platforms as prototypical implementations was conducted. These reviews of literature will be used as foundations for models developed in the next chapter as well as requirements collection for the prototypical implementation of a mobile multimodal serious games analytics system.

# 4 LeGUC Serious Games Analytics Model

In this chapter, *the LeGUC model* (Learning-Gaming-Using-Context) is introduced as a general theoretical foundation defining dimensions of Serious Games Evaluation. This was needed to help in defining a concept for linking serious games log data with affective and cognitive states, user behavior and context data This part will be introduced in the second section and is called the *Reasons and Responses Model*. The models are initially based on a literature review on related models in the fields Serious Games, Affective Learning, Context-aware Mobile Learning and User Experience. Then it was put into use in several evaluation processes which helped expand and refine it.

## 4.1 Model: Dimensions of Serious Games Evaluation

In Section 3.4, a comparison of available theoretical models covering dimensions of Serious Games evaluation was conducted in Table 3.5. Most of these presented frameworks do not:

- distinguish which aspects of the game need real testers and which can be evaluated using only heuristics and inspection
- distinguish which evaluation aspects will differ from one user to the other and which are independent from user/player attributes
- consider the usability dimension for evaluation
- address the issue of alignment between different dimensions
- help map observed interaction indicators to the evaluation of Serious Game elements

The need for a model covering all dimensions found in literature as well as being more precisely descriptive of the nature of interactions between the play-learner and the serious game is the motivation behind our work. The model presented in this section is based on all previous models presented in Table 3.5 and is an attempt to fill the identified gap as a foundation for the more in-depth model aimed at multimodal data which will be presented in Section 4.2. Although aspects from all three dimensions discussed in the previous chapter are strongly interrelated, we have chosen to divide them into the three categories (learning, gaming, using) ignoring their interdependencies, see Table 4.1.

In addition, according to the available literature (cited in Table 4.1), we have differentiated the following categories:

- **static product aspects:** aspects of the product to be tested (the serious game) which can be determined by mere inspection without the need for user testing as they are user-independent
- **dynamic product aspects:** aspects of the product (serious game) that need interaction from users (play-learners) to be evaluated and are thus user-dependent
- **dynamic user aspects:** aspects of the user (play-learner) that can be determined through and occur during the interaction with the product (serious game)
- **static user aspects:** traits of the user (play-learner) that are static but may still be determined through the interaction with the product (serious game)

Thus in summary, we define aspects which can be collected without the need for user interaction with the product as static whereas dynamic aspects need interaction to be observed or calculated. The aspects collected from literature are summarized in Table 4.1. The second (dynamic Serious Game aspects) and the third category (dynamic play-learner aspects) are in the focus of this work as they are the ones which play a crucial role in studies evaluating Serious Games with users. These two categories are interrelated as aspects in the second category once determined can help in evaluating aspects in the third category.

**Table 4.1:** Aspects to be Evaluated in a Serious Game.

| | | Learning Aspects | | Gaming Aspects | | Using Aspects | |
|---|---|---|---|---|---|---|---|
| | | **Learner Model** | | **Player Model** | | **User Model** | |
| | | **Affect & Cognition** | **Performance** | **Affect & Cognition** | **Performance** | **Affect & Cognition** | **Performance** |
| **Play-Learner** | **Static** | cognitive traits, preferences, style, interests [115, 70, 181, 102] | prior knowledge, skills, abilities [70] | player type, affective model [24, 81, 101, 86] | former gaming experience [140, 105] | demographics, design preferences, attitude towards SG [142, 158, 102] | former experience with SG [257] |
| | **Dynamic** | interest, workload, attention, distraction, drowsiness, confusion, recalling, reflection, confidence, pride.. [103, 277, 220] | response delay, response accuracy, numOfAttempts, self-correction, learning-related actions [93, 230, 247, 139] | surprise, engagement, excitement, fun, immersion, anxiety, relief, .. [123, 148, 59, 162, 79] | time to complete, progress, score, levels, game variables, paths, assets, game-related actions [50, 140, 49] | confusion, frustration, anger, pleasure, satisfaction, .. [18, 182, 192] | click location, click speed, inactivity mouse movements, navigation, time of access, tools usage [28, 142, 252] |
| **Serious Game** | **Dynamic** | pedagogical agent, instructions, feedback, reward, helping focus [27, 245, 269, 280, 271] | info chunking, scaffolding, difficulty level, pedagogy, connections to authentic uses, stimulation of further inquiry [286, 103, 96, 100, 245] | immersion elements, motivating elements, clear goals and subgoals [123, 58, 79, 148, 50] | challenge elements, complexity, replayability [123, 58, 79, 148, 50] | clarity, efficiency, meaningfulness, recognizability, context-Sensitivity [27, 6, 286, 59, 236] | responsiveness, ease of use, intuitiveness, learnability, no irreversible errors [27, 6, 286, 59, 236] |
| | **Static** | relevant theme, consistency of approach [27, 245] | scope and depth, domain relevance, activities diversity, self-assessment [130, 245, 103] | story, imagination, positive role models, visible progress [245, 286, 59] | diversity, degree of freedom and self-expression, social elements [286, 59] | quality, consistency, helps you know where you are [236, 27, 6, 286, 59] | option to skip content, clear exists and main menu everywhere, saves states & info, provides help [27, 6, 286, 59] |
| | | **Presentation** | **Mechanics** | **Presentation** | **Mechanics** | **Presentation** | **Mechanics** |
| | | **Learning Design** | | **Game Design** | | **Interaction Design** | |

For example, determining cognitive states like cognitive load and effort can make statements about the size of learning modules and the difficulty level. This will be examined in more details in Section 4.2.

The resulting theoretical framework for Serious Games Evaluation can be seen in Figure 4.1. In this construct the focus is on Single-Player educational games. Within each of the three dimensions (Learning, Playing and Using) we distinguish, as in Table 3.5, Experience (Affect and Cognition) and Performance on the user side and Presentation and Mechanics on the product design side. The reason we put affect and cognition together is that many researchers have found it difficult to draw a concrete separating line between the two as the way we feel mostly cannot be separated from the way we think.

As the literature review (see Table 3.5) revealed the importance of coherence in design, we have as a parameter the alignment between the three dimensions. This means that it is important when designing the gaming elements of a Serious Game to take into account that these elements should at the same time be supporting the learning design elements and not disturbing learning, for example. The same applies to the relations between all design dimensions which are indicated by arrows in Figure 4.1.

As laid out in Section 2.5, the context in which the interaction takes place also plays an important role in the evaluation process [261, 13, 109]. Aspects related to context can be:

- Time, location
- Environment conditions (illumination, noise, weather, location)
- Device conditions (orientation, shaking, speed)
- User conditions (activity, social interactions - also mood, stress,..-)

When using this framework for design and evaluation, it is important to note that, like designing any product which needs to satisfy conditions in different dimensions, trade-offs between the different dimensions need to be undertaken, especially when conflicts of interest arise. Thus, it might be useful to use a model such as the PLU Model (with the three dimensions Playing, Learning and Using) [161] to determine appropriate design and evaluation decisions, e.g. giving each dimension a percentage according to importance.

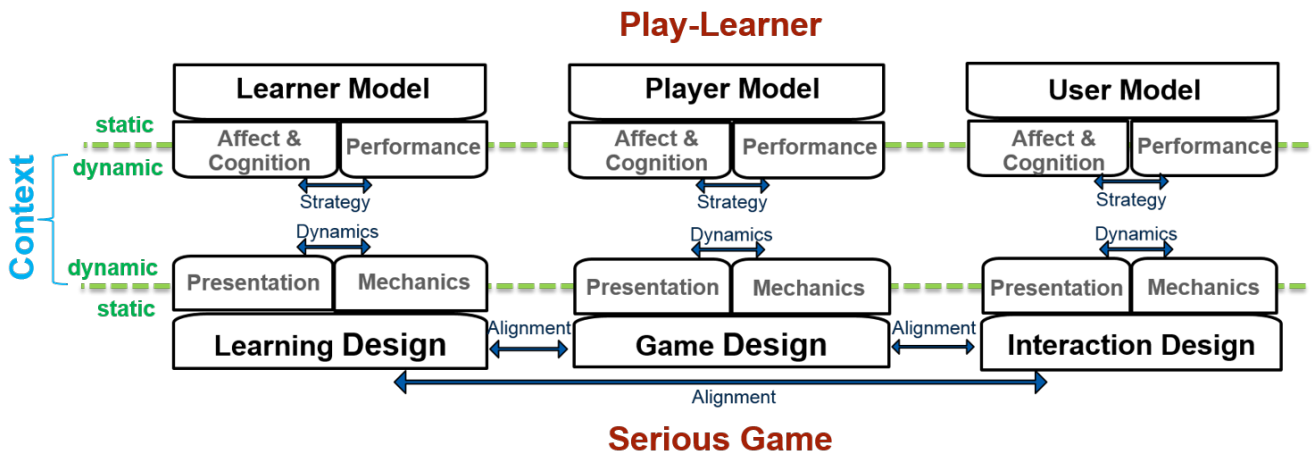



**Figure 4.1:** Model: Dimensions of Serious Games Evaluation.

## 4.2 Augmenting Event Logs with Multimodal Data: The Reasons and Responses Model

After determining what is to be evaluated in a Serious Game (see RQ1), there was a need to determine how these aspects will be measured and where multimodal data can add more value to the evaluation (to answer RQ2). This will be examined in more detail in this section.

### 4.2.1 Model: Reasons and Responses

By looking at Figure 4.1, one can notice a core element which repeats in all the three dimensions. This core element consists of the dynamic evaluation aspects: Presentation, Mechanics, Affect/Cognition and Performance. These four aspect groups constitute the main elements contributing to an evaluation process with users (after examining all static aspects). Thus, to determine where there is a need for multimodal data inside the evaluation process of Serious Games, one needs to take a closer look at the relations between these four aspect groups. This is illustrated in Figure 4.2 which shows the Reasons and Responses (R & R) Model, a core contribution of this research. The first step is to examine which aspects can be captured using event logs. Event logs typically capture actions of the user (user-originated events) inside the game environment, e.g. a certain move or clicking on a button, in addition to game events triggered which the user sees as part of the presenation, e.g. start of a new scene or positive feedback (system-originated events). Some studies try to analyze these events to infer the users strategy or motives but they mostly find ambiguities in the reasons behind actions if only logging is inspected. This is because the reasons behind a user action can lie in the mechanics of the game, the experience of the player which can be expressed through affective and cognitive states or they can be independent from the game and the user as they can lie in the context. In this case, pure logging events would be ambiguous and the researcher or the system will need more information to make any decision. Affect and Cognition can also be studied separately as a response to game events (presentation) to examine effects induced on the experience by certain game events. Examples of these ambiguities will be discussed in the following section.

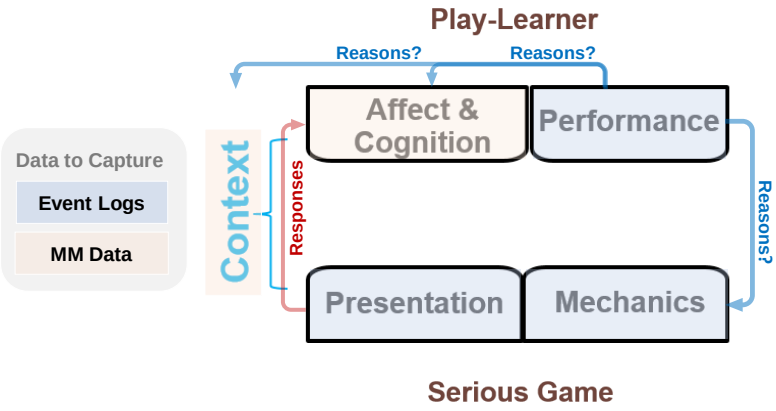



**Figure 4.2:** Model: Reasons and Responses.

### 4.2.2 Application of the Reasons and Responses Model on Evaluation Dimensions

In this section the Reasons and Responses (R & R) Model will be applied to the aspects described in Table 4.1 to illustrate why there is a need for multimodal data for determining/disambiguating reasons behind actions in a Serious Game evaluation. Applying the model on the Learning Dimension as in Figure 4.3, if event logs have captured an interesting behavior, e.g. that the learner has paused before answering a question, then this may be due to several reasons: The reason could lie in the learning design of the game as the learner may have a problem in understanding and thus did not benefit as expected. However, the reason might also lie in the experience of the presentation if the instructions for answering the question are confusing. But both inferences may be wrong as the reason might simply be that the user has received a phone call before answering the question, a factor which is related to the context. So this is an example of an ambiguity in analyzing event logs where multimodal data can help give more

insight. An example of determining a response to events on the learning design side where pure event logs may not be enough is determining the effect of positive feedback. This can be detected by looking for emotions such as pride which are expected to be induced by positive feedback. In Figure 4.3, reasons are indicated by blue arrows while responses are illustrated by red arrows.

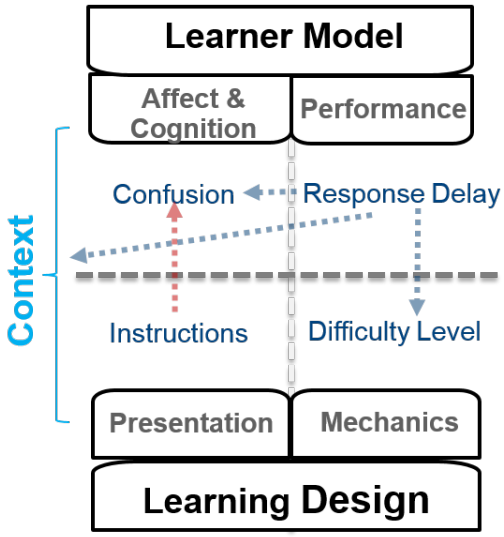



**Figure 4.3:** R & R Model Applied on Learning Dimension.

On the gaming dimension (see Figure 4.4), skipping a subgame might mean something about the replayability of the subgame or not feeling excited enough in this subgame if its design lacks motivational elements, for example. Another reason might be that the subgame is not suitable for the current context, e.g. instructions cannot be heard in a noisy environment. To examine responses to game events, the researcher may want to check if a user is immersed in the game during a story peak which is not always directly inferable by analyzing event logs alone.

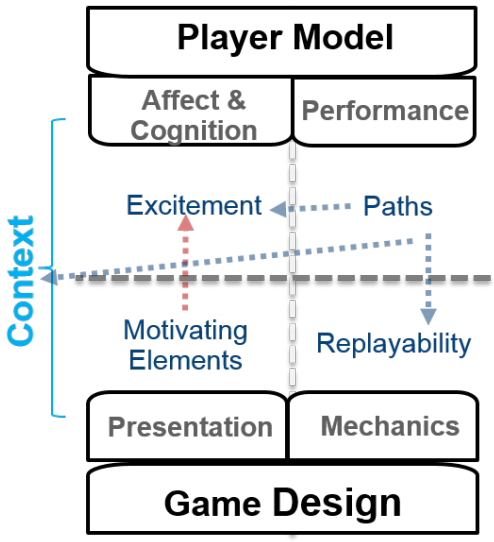



**Figure 4.4:** R & R Model Applied on Gaming Dimension.

On the using dimension, a common trace is the use of the help function in a certain scene. This can be used to assess proper interaction design, but the user might actually be confused with something else in the game or simply explaining to a friend how to use the application (see Figure 4.5). So these were examples on the three dimensions but the model can also be applied across dimensions. An action in the learning domain (e.g. delay in answering a question) can be a result of a problem in the using domain

for example (non-intuitive screen design). This can be used to detect misalignments between different dimensions [96].

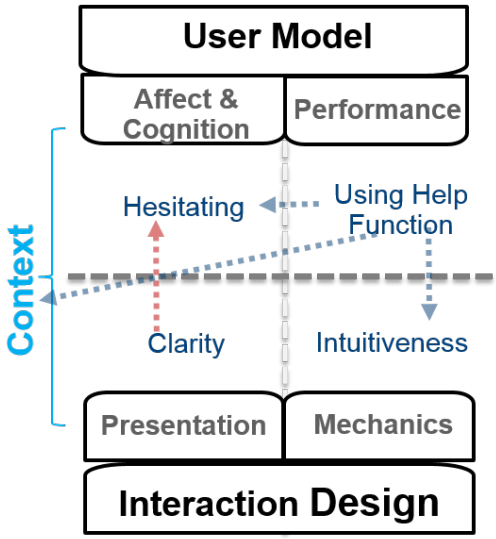



**Figure 4.5:** R & R Model Applied on Using Dimension.

## 4.3 Choosing Suitable Data

Different types of data can be used for capturing gameplay in serious games. Logging is the most traditional and widely-used method for recording actions. This includes time-stamped actions, game variables, world coordinates, assets and paths [141]. Defining the grain-size and the type of logging data to be collected is an important decision as it is expensive and time-consuming to collect all possible data and then search through it to find the needed information. The best unit size with the smallest meaningful action is called "operant action" by Schell [226]. Deciding to collect either *generic or game-specific* data [240] depends on the primary goal of the evaluation. While generic traces are not tailored to the specific game design features but are suited for comparing games or game modules, game-specific analytics take into account differences in designs and requirements resulting in the need to be defined by game designers or educators according to the aspects they need to uncover. Depending on whether quantity or quality is desired, *intensive or extensive data* can be collected [209]. Considering the criteria types differentiated in Table 4.1, extensive data can be used, for instance, for dynamic (user-dependent) product aspects as the focus would be more on game elements whereas intensive data could be more suitable for dynamic user aspects as the focus would be on a deeper analysis of the users themselves. In addition to logging, researchers or designers can have access to a replay of player actions inside a Serious Game by using a system which reproduces a game session from its logging data alongside updates of internal game states combining the benefit of both video analysis and logging [96].

As discussed in Chapter 4.2, logging alone is typically not enough for capturing aspects like affect, cognition and context. For deteriming affective and cognitive states like arousal, attention, focus, and effort, visual and behavioral cues like facial expressions, gestures, posture, head motion, eye tracking, blinks and shaking can be used and mapped to specific design elements, game experience or educational achievement [232]. Through replaying and/or analyzing audio e.g. detecting laughter or verbal expressions of affect or analyzing/categorizing think-Aloud expressions, it is also possible to detect cognitive and affective states [132]. A very popular method for measuring affect and cognition like cognitive load, stress and arousal in the evaluation of Serious Games is using physiological sensors and analyzing bio-signals like Skin Conductance, HeartRate, EEG/EMG/ECG, etc. [182]. As for context, visual fea-

tures determining illumination or distractions, for instance, as well as audio features determining noise, distractions and conversations, can be used.

Mobile sensors can be used to determine *environment conditions* like illumination, noise, weather and location, *device conditions* like orientation and shaking, as well as information about the *user* like his/her activity, social Interactions, stress, emotion and speed using front-faced camera, microphone, GPS and accelerometer and about his/her *interaction* such as touch pressure and typing speed.

When attempting to choose a suitable modality for a certain analysis task, there are many criteria to be considered. The first consideration tackles the cost/benefit ratio as some modalities are very expensive and difficult to deploy but may deliver poor results in the measurement of a particular aspect. Another important factor to differentiate modalities is how much invasive they are, which may put limitations based on the evaluation settings and the reliability of the results the researcher wishes to obtain. Availability also plays an important role as some modalities, while more accurate than others, might be difficult to obtain [36], for instance some recognition techniques might be prohibited in the classroom. Thus, the decision of choosing suitable data should be studied very carefully before making decisions for an evaluation process [25, 37].

In certain contexts, the detection of some information is not possible when using only one modality for recognition but only by integrating multiple modalities. In [283], only multimodal analysis succeeded in evaluating flow. As our human interactions are also multimodal in nature, more and more researchers are moving towards multimodal assessment especially when studying emotions [107]. For instance, a teacher in class will perform better when considering not only students verbal expressions but also their visual behaviors and using messaging for discussions is not as effective and powerful as face-to-face conversations. Each bit of information adds richness to the interpretation of data.

A summary of possible multimodal data which can be used for the evaluation of Serious Games and their different corresponding purposes is presented in Figure 4.6.

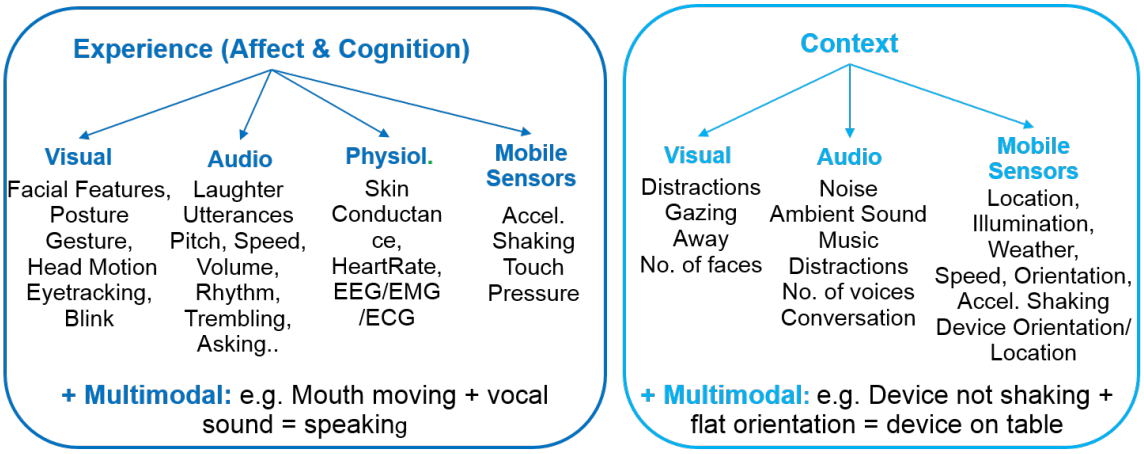



**Figure 4.6:** Possible Multimodal Data.

## 4.4 From Event Logs to Multimodal Data

The Reasons and Responses Model discussed in Sections 4.1-4.3 illustrates the potential use of multimodal data for uncovering strategies and experiences of users during evaluation studies of Serious Games. This section will illustrate how a researcher can decide when to look at additional multimodal data. As event logs can be automatically captured by analytics components integrated into the game, it would be useful for the researcher to have an automated aid to spot “interesting” behavior which might require looking at additional data for investigation. Figure 4.7 illustrates such a logging system which takes into account the possible need for additional multimodal data.

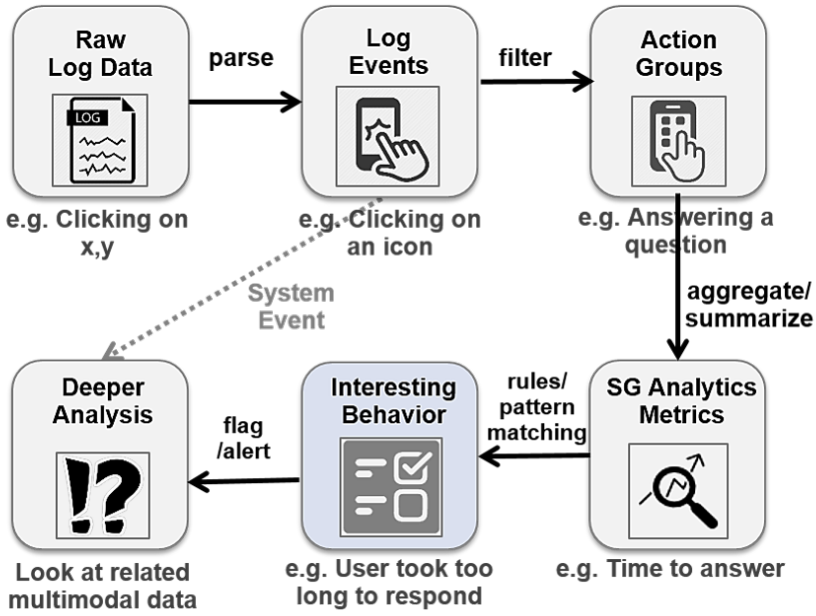



**Figure 4.7:** From Event Logs to Multimodal Data.

Raw log data such as mouse click positions are first parsed to detect events like clicking on an icon and then filtered into trees of action logs which can contain other action logs, e.g. drag and drop or answering a question. Metrics are calculated from these events which can then be used to detect ambiguous behaviors and alert the researcher and/or automatically analyze other multimodal data. The process of detecting such a behavior in a scene-based Serious Game can be done as follows:

The first step is to segment data according to events like new scenes or updates in the internal model for determining, for example, the frequency of mouse clicks in a scene. To do this for a particular user differences between scenes need to be considered. To account for this, a weight can be assigned, either initially determined by a pilot study or defined by the researcher. To cancel the effect of scene differences, this weight should be inverse proportional to the average number of clicks collectively calculated from all users so far to detect an interesting behavior of this particular user. The normalized average of a certain metric (for this example it is the number of clicks in the scene) can also be used to distinguish interesting user behavior for a certain scene across all users and so it can be used to identify common paths and interaction patterns of this game.

So this is defined here as the extensive indicator. Using this for each metric and making a sum of products of all intensive metrics for the current user across all scenes (with their respective weights) for a certain evaluation aspect, a value can be obtained for the evaluation aspect for this particular user. We call this the intensive indicator. So this may point the researcher towards a particular inference from the log events alone which may also require further investigation in multimodal data.

## 4.5 LeGUC States in Serious Games Evaluation Sessions

Most studies detecting affective and cognitive states within serious games interaction sessions were limited to a small number of traditional emotions (see Table 4.2) which do not cover all aspects related to the three dimensions learning, gaming and using in addition to the context dimension, which are, as discussed in this chapter, crucial for understanding learner experience in serious games [192]. A more general approach was thus needed to include a reasonable set of states worth capturing during a serious games evaluation session in the field.

This should consider aspects related to affect, cognition, behavior and context which are possible to encounter during an educational game session and related to the dimensions: learning, gaming, using

and context. To do this, it will need to take into account measures originally defined for pure learning environments, others defined for games as well as measures defined for user experience evaluation. The focus of the study is not how to recognize these features using automated methods nor which modalities are best suitable for capturing the data but rather define which aspects need to be considered to gain a better understanding of the full experience.

As event logs typically capture interactions with the game environment, it needs in many cases to be enriched by user reactions and other context data to be more effectively analyzed. Consider a user being inactive for a remarkable time period inside a serious gaming session. The log data will capture the time elapsed before an interaction occurs but more data will be needed to disambiguate the reasons behind this behavior to properly determine the resulting design decision.

The reasons for such a response delay can lie on the learning dimension if the user is found to be reflecting on the answer to a question or on the gaming dimension if the user is reflecting on the strategy to make the next game move, for example. Furthermore, the problem might be related to usability if the user is found to be confused on how to proceed with the navigation, for example.

However, the reason might also be exterior to the whole gaming experience, as it might lie in the context: the user might have been distracted by an event in the environment which happens frequently in real-life situations or when testing in the wild. All these factors need to be taken into account for an ideal evaluation. For this, we will assume an observer watching the testing session and being able to take note of states related to these dimensions which can have influence on the experience at the moment they occur. This would be the ideal situation for capturing all relevant data which, in real testing scenarios, however, will naturally suffer from many limitations. Nevertheless, it can serve as the ultimate goal of recording the experience, whether doing this manually or automatically. Table 4.2 presents parameters of Affect, Cognition, Context and Behavior considered in different models investigating data in learning environments.

After reviewing the different parameters considered in literature, it was found that the best approach would be to carry out a study to determine which of the factors present in literature will be elicited during playtesting sessions of Serious Games carried out in the field. This would help decide which of the above listed models more precisely fits the evaluation of (mobile) educational games in natural environments on each of the three categories: Affective and Cognitive States, Context and Behavior.

Based on the results which will be discussed in the Evaluation chapter, the chosen states are depicted in Figure 4.8.

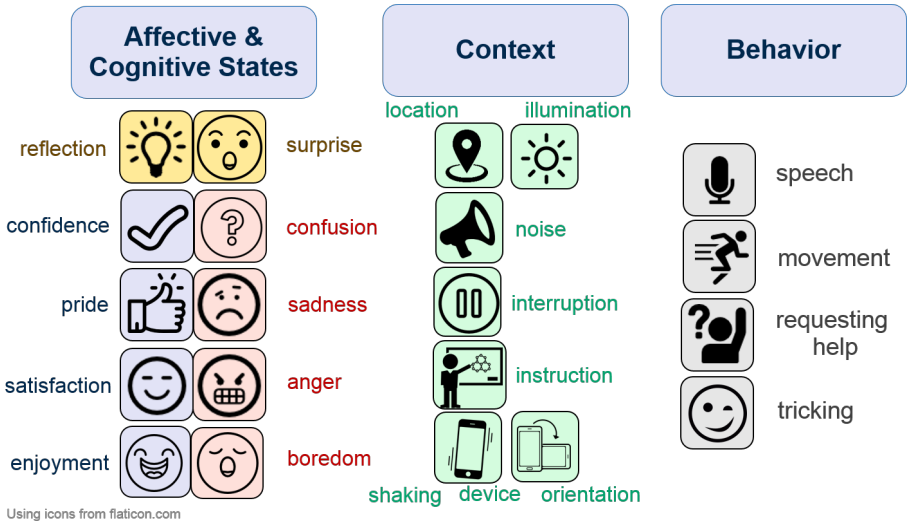



**Figure 4.8:** LeGUC Features for Evaluating Experience in Serious Games Testing Sessions.

**Table 4.2:** Evaluation States Identified in Literature to Occur in Interaction Environments.

| Citation | Environment | Affective and Cognitive | Context | Behavior |
|---|---|---|---|---|
| [194] | Online learning | frustration, anxiety, shame, excitement, pride | | |
| [51, 53] | Educational games | joy, regret, admiration, reproach, pride, shame | | |
| [19, 18] | Intelligent Tutoring Systems | boredom, confusion, delight, engaged concentration, frustration, surprise, neutral | | on-task, on-task conversation, off-task Conversation, off-task solitary behavior, inactivity, gaming the system |
| [79] | Game-Based Learning | interest, enjoyment, task involvement, confidence | | |
| [234] | E-learning | interest, engagement, confusion, frustration, boredom, hopefulness | | |
| [201, 203, 204] | Academic settings | enjoyment, hope, pride, relief, anger, anxiety, shame, hopelessness, boredom | | |
| [91, 52] | Educational games | boredom, shame, frustration, confusion, disappointment, surprise, neutral, curiosity, engaged concentration, delight, excitement, confidence, pride | | |
| [29] | Mobile learning | | location, time, weather, temperature, noise, lighting, day, movement, device capabilities | |
| [284] | Mobile learning | motivation | noise, busyness of the environment, temperature | |
| [85] | Mobile Learning | | people, hardware and software capabilities, place, time, noise, illumination | |
| [67] | Mobile learning | enthusiasm, boredom, happiness, sadness , satisfaction, calmness, anger, anxiety, frustration, fear, confusion, hope, pessimism, expectancy , astonishment, sympathy, disgust, hate, pride, shame | time, location, terrain, weather, neighbors mobility, device capabilities | |

## 4.6 Conclusion

The need for understanding serious play experience has led to more and more research being conducted to use multimodal methods in the evaluation of serious games. However, available theoretical evaluation frameworks for Serious Games do not consider the use of multimodal data despite the increasing number of studies using them and despite their many benefits discussed in this chapter. After an overview over available frameworks for the evaluation of Serious Games was given in the last chapter, the Reasons and Responses Model was presented, with the aim of filling a gap in theoretical foundations for Serious Games evaluation by focusing on the value of adding multimodal data to event logs. First, dimensions of evaluation aspects were defined based on a literature review. Then it was examined which role multimodal data can play in measuring these aspects, specifically in determining reasons behind users' logged gameplay actions and their responses to game events using the proposed model. Finally, practical examples of using this model for combining logging with multimodal data in evaluation were discussed. Establishing this framework, which was also published in [238, 246] will be applied in developing a

multimodal Serious Game evaluation tool as a proof-of-concept which will be the next step in the current research. It is answering parts of Research Questions 1-3 as it defines parameters for a serious games evaluation (Research Question 1), defines a hypothesis of why and when multimodal data is needed to interpret log events in Serious Games Evaluation (Research Question 2) and defines different types of multimodal data needed in evaluations and when they can be used (Research Question 3).

# 5 StoryPlay Multimodal: A Research Platform for the Mutimodal Evaluation of Serious Games

This chapter presents the design and development of *StoryPlay Multimodal*, a mobile multimodal analytics platform for the evaluation of Serious Games. It is intended to serve researchers, teachers and educational game developers as a means to assess their Serious Game Design. This is done by capturing, pre-processing, synchronizing and visualizing multimodal serious games analytics and mobile sensor data from playtesting sessions. By linking log data with multimodal data, it is possible to uncover relations between design elements, gameplay interactions, context parameters and affective and cognitive states. This is crucial for gaining full insight into a session, even if not present with the player at the same location. After discussing design requirements, the architecture of the software, the different modules, additional features, implementation challenges and solutions are presented. Parts of this chapter are published in [237, 239].

## 5.1 Design Requirements for Mobile Multimodal Serious Games Analytics

The design goal of StoryPlay Multimodal (StoryPlayMM) is creating a non-invasive Serious Games research/evaluation tool supporting remote, asynchronous observational evaluation of mobile serious games. Main requirements for the framework were determined from literature and software review (see tools in Tables 3.6 and 3.7 as well as recommendations of Serious Games researchers.

### 5.1.1 Underlying Architecture

The StoryTec Authoring Environment is built upon an internal model considering updates in the learner and player model during play in addition to the storyline [166]. The original StoryPlay rapid prototyping tool is based on the same story engine (see Figure 5.1) and also displays updates in the internal models [217]. This data is gathered based on the Narrative Game-Based Learning Object (NGLOB) Model [81]. The player used for running games created with StoryTec, called StoryPublish, allows running the game interface on different platforms [242]. The story structure is formatted using the xml-based model description language ICML [81] and communicated between the authoring tool and the Story Engine. This same information is used for reconstructing the sessions in StoryPlay using StoryPublish.

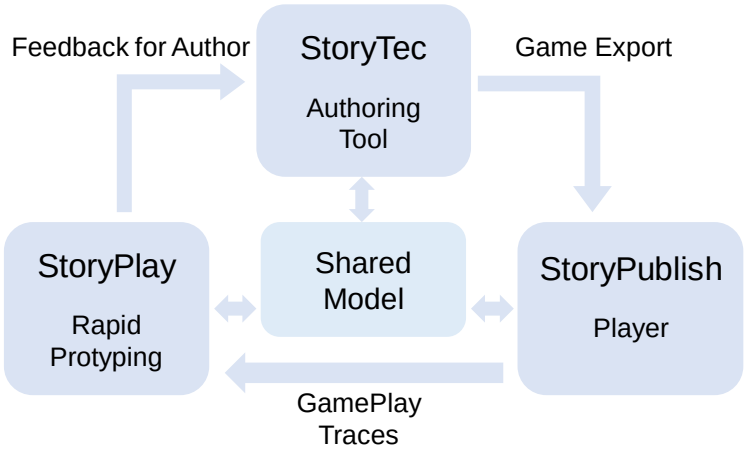



**Figure 5.1:** StoryTec Architecture.

### 5.1.2 Design Goals and Requirements

The goal of this work is to extend StoryPlay to support multimodal data and link it with event logging and internal model updates while minimizing invasiveness. This would be helpful for authors/researchers/instructors (here StoryTec users) to offer them a ready means for getting feedback to improve the design of their serious game. To minimize obtrusiveness, the prototypical implementation is developed and tested on smartphones where the sensing mechanism is regarded to be far less obtrusive than sensors which are worn on the body or fixed inside labs. This also allows carrying out evaluations by play-testers worldwide without having to be present in the same place.

Using mobile sensors also gives the possibility of providing an insight into a wider set of context parameters as context plays an integral role in playtesting nowadays due to its influence on experience. The hypothesis is that an easy and goal-oriented navigation through multimodal data would help researchers disambiguate ambiguous actions in the event log. The ultimate goal is to allow researchers to better understand relationships among data and provide them with additional information from natural settings. For instance, finding out if a pause in gameplay activity is due to experiencing frustration, reflecting on playing strategy or learning content or getting distracted by the environment can be important for evaluation (Other examples of such ambiguities can be found in the applications of Reasons and Responses Model described in Section 4.2.2).

Relying on data automatically captured during the gameplay experience should help make evaluation studies more objective and time-efficient than subjective observation and self-reports for uncovering aspects related to emotion and cognition. Combined with logging, this should help identifying advantages and problems with specific game elements with regard to fun, usability and effectiveness and how to improve the Serious Game at hand thus helping advance Serious Games research.

The requirements for the development of StoryPlayMM arose from a real practical need identified by the Serious Games research group in various research projects and not just on theoretical assumptions. As mentioned earlier, it was a development objective to integrate the most useful features from different platforms into one platform tailored for Serious Games. Features considered for design are gathered from available research and commercial tools discussed in Section 3.5 and adjusted to be used with scene-based Serious Games authored with StoryTec. The tool should be usable without prior programming skills to make it usable for all StoryTec target users. It should also have potential for integrating different recognition modules to act on the analysis of the captured data.

As in the PlaytestCloud tester app mentioned earlier, StoryPlayMM has a similar approach in recording log files on mobile devices, as well as front-facing camera video and/or microphone recordings. Audio recordings can be used for think-aloud, laughter detection etc. - and detecting the environment background noise.

The use of a session video view with adjustable speed is desirable for investigating in-game action in parallel with their responses [187]. In addition, allowing the user to quickly skip through a video to jump to a certain event or a certain reaction by clicking on this particular interest point was found to be a very useful feature [187]. In StoryPlayMM, this is achieved not by recording screen video which would be big in size along with the front-facing camera video and not easy to navigate to a certain event. It is achieved by a session replay tool which reconstructs gameplay from the log files using the game engine itsself, as will be discussed in next sections. A replay of the whole gameplay session is reconstructed from the logged information by embedding the StoryPublish player in addition to enabling adjustable speed and interactive navigation of the replay based on main events. The navigation feature helps in speeding up the process of evaluation as researchers can directly navigate to the specific event of interest without having to watch the whole session.

A close coupling with StoryTec authoring (.icml) files allows saving much space in the replay component files. This tight integration allows accessing and showing internal state information along with the interface itself to assist in correlating game activity with leaner experience. With this replay, there is no need for large and difficult to navigate files containing screen recordings.

The feature of recording and replaying face expressions of game players is presented as having many advantages for evaluation: It is an unobtrusive way of observing players' engagement and involuntary reactions during playtesting [63, 61]. Recognition and Analysis of these expressions can also be applied on the recordings as a quantitative assessment method [262].

For the logging, main recorded actions from the player side and events from the system side need to be distinguished and represented with the possibility to jump to related multimodal data where needed. In addition, mobile sensor information like illumination and movement can help determine context as discussed in Section 4.3. Location sensors have been used in some multimodal interaction analysis tools supporting mobile deployment to track the location of the testers during a remote testing session [77]. This is useful when having playtesters from different places and a need to gather location data and was thus also included in StoryPlay Multimodal.

The synchronization and interactive navigation of multimodal data is applied in many multimodal data analysis tools like ChronoViz [75], Tatiana [66], Digital Replay System [38], Noldus Observer [290] and Mangold Interact (mangold-international.com). It enables researchers to jump to the point of interest and see all related multimodal data run simultaneously next to each other which saves time and effort of analyzing and annotating qualitative data.

The feature of filtering out uninteresting video frames using information from low power mobile sensors like the illumination sensor was described in [94]. They also investigated predicting whether a frame contains faces using the accelerometer and gyroscope sensors. These features have been included in StoryPlay as well and this was one of the reasons why mobile sensor data is used in the framework. The other reason is that it also provides data on how the player is holding the device which is also important for UX testing.

Built-in measures of self-reporting, testing instructions and after-game survey can be integrated to investigate correlations with observed interactions and reactions. Eye-tracking and Physiological data are not directly included in design considerations, but can be extended if there is a non-obtrusive way of monitoring using built-in mobile sensors. It is possible for users and developers to select data sources to avoid privacy issues. Player profiles can be created or, if permitted, can be collected from the phone.

Session summarizations include aggregations and a future feature would be detecting and highlighting relevant data such as correlations, repetitions and significant events and sequences. Detecting event sequences may be helpful for comparing the input stream with some target sequence (e.g. of an expert player) or emphasizing certain pre-defined player behavior.

Aggregations can be done across users for extensive data (e.g. scenes where most users had a sudden movement or smile or laughter etc.) or across modalities for one and the same user for intensive data (when more than one data source has significant change at the same time) [240]. Abstracting low-level details for investigators helps speed up the evaluation process and aggregations can help in rating subgames as well as identifying common paths and interaction patterns. The hybrid approach where analysis is not fully automated and not fully manual is chosen for our purpose as it is difficult to fully rely on automatic recognition where experts can more efficiently get better results [255], especially because of the heterogeneity of Serious Games. These experts, however, still need the pre-processing to save time and have objective measures. There is a potential for supporting the integration of off-the-shelf recognition modules, e.g. for facial features. A way for supporting user-defined annotations should also be included.

In summary, the most important features for the platform can be summarized into the following and described in [238]:

- supporting the dimensions of learning, gaming and interaction
- recording AND playback of multimodal data
- logging or screen recording of interaction
- supporting different modalities like video, audio, eye-tracking, physiological and mobile sensors
- synchronization of multimodal data
- ituitive interactive navigation

- filtering and preprocessing of data
- a means of annotating data
- data analytics and visualizations

An overview of the architecture and design goals is given in Figure 5.2.

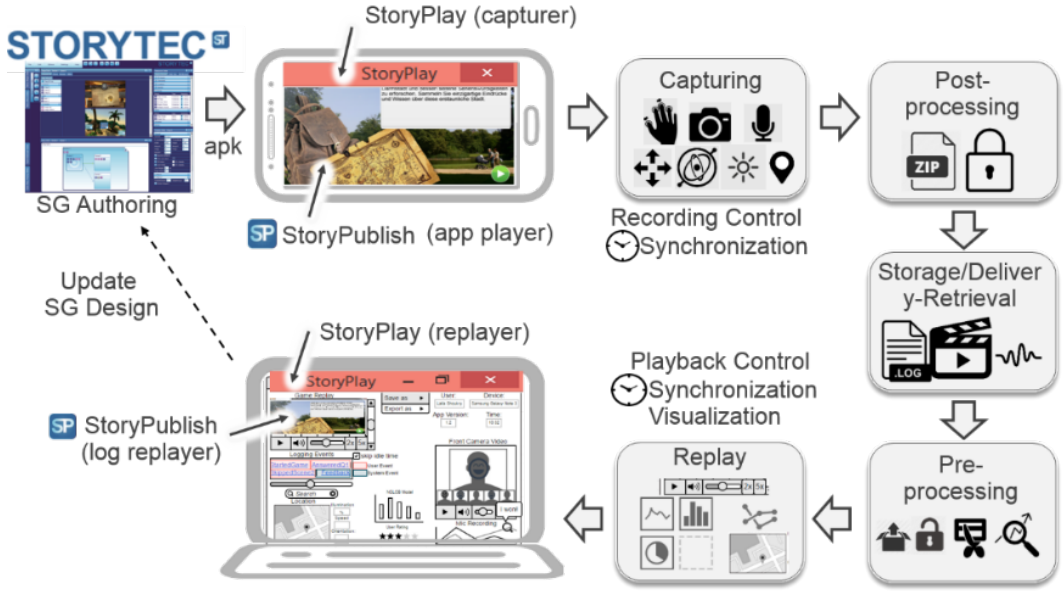



**Figure 5.2:** Multimodal Evaluation Platform Design Architecture [237].

Two different modules are needed for realizing this tool: a capturing app which is run with the application to capture logs and multimodal data (here a mobile app as a prototypical implementation of a naturalistic testing session in the field) and the main desktop analysis tool for replaying and analysing data.

## 5.2 Design and Implementation

StoryPlay Multimodal platform design consists of two main modules: The **Capturer** app and the desktop **Replayer** component. A mockup of the tool with the features described in the previous section is depicted in Figure 5.2. In addition to those two components, an **Observer** app was developed for testing sessions where an observer is present to help him record his/her session observations in real-time, which are then also shown later according to their timestamps in the session replay. In this section the basic features implemented in the different modules are described in more detail. Some interfaces of the modules are shown in Figures 5.3-5.7.

### 5.2.1 Capturer App

The Capturer mobile application was integrated into mobile educational games exported with StoryTec [166, 242], a Serious Games Authoring Environment used for creating scene-based educational games based on the NGLOB (Narrative Game-Based Learning Object) Model [81]. To combine the advantages of event data and the more rich observational data, it was a main requirement of the project to provide the investigator with a way to review all segments of the session quickly without having to manually search through the entire data as well as get help and summary statistics. All events can be used to navigate in the session by skipping to the timestamp where this event occurred. Interesting parts of the video are also highlighted as we will see later. Ways of seeing where a new scene started and which

notes have been provided by the observer on a particular scene. Also mobile sensor data are visualized and their timelines are also synchronized with the timelines of the events, observer events, game and video, all playing together.

Based on design requirements identified in [238], the following are the major features supported on the capturer app:

- synchronous recording of video (from mobile front-facing camera), event logs and mobile sensor events with flexible user configuration options
- integration inside StoryPublish Android software (used for running StoryTec games on different platforms), run on different Android devices
- unobtrusive, not negatively affecting game experience
- storing and retrieving user profiles on device
- avoiding privacy, storage and bandwidth issues by giving user full control over sensor activation and over what and when to store and/or upload to the server
- interface usable without programming or special background
- providing a game rating option after gameplay


**Figure 5.3:** Mobile Capturer Component Screens before and after GamePlay.

The capturer app (some screenshots are shown in Figure 5.3) is an extension around the mobile version of the StoryPublish engine which wraps the game to be tested after it is exported to Android from StoryTec Authoring Software. This wrapper is implemented in haxe based on kha engine. The capturer contains the options for recording user, log and multimodal data during interaction with the

mobile edugame. When running the app, the player first enters his google drive log-in data if he wishes, if not it is set to the default firebase server [1] which offers cloud storage for app data. After setting the user profile for testing (name, age, gender and class), the user chooses which sensor data are recorded and will have the option at the end to give permission on what is sent, what is saved on the device and what is deleted. S/he can check any or all of the following sensors: location, camera, accelerometer, gyroscope, proximity and illumination using icons and descriptions for the different sensors. S/he can also choose the sample rate of some sensors to be normal, high or very high using a slidebar. In the case that the researcher is co-located with the tester, these options can be chosen by the researcher. Also if it is a child testing remotely, this part can be done by his parent. During gameplay, the sensor data specified is logged and if given permission, the front-facing camera captures a video of the participant's face during play. The recording part is implemented mostly in java. At the end of the gaming session, a rating scene is shown where the player rates his experience.

Log files saved on the mobile phone (to be sent later to the server) contain timestamped logged events from the game like clicks, variable changes and scene names in addition to events needed for synchronization and statistics like timers as well as events recording sensor changes. In addition, some events were added for mobile gaming like game pause and game resume to account for interruptions in game play.

### 5.2.2 Observer App

The observer app was initially not part of the design but was then found to be important for several reasons: First, it was found to be a good addition to the evaluation suite, as it can be used in observational studies where the researcher is co-located. Secondly, it can help in training a machine learning algorithm for extracting features from multimodal data by serving as a ground truth. Lastly it helps investigating relationships between log events and affective, cognitive and context states assuming that we have already extracted the given features from multimodal data which can be a great help to advance Serious Games research.

Taking notes during a session by the observer can be so time consuming that an important observation can be missed while the observer is still writing. So the main goal of the interface is to make it very easy to record observations with just one click during the usually fast and unpredictable playtesting session. The design and choice of the observation recording buttons and types is based on the LeGUC Features for Evaluating Experience in Serious Games Playtesting described in [246] and depicted in Figure 4.8.

To the best of our knowledge, this is the first observation app considering features related to Serious Games Evaluation. The current design after several iterations and tests can be seen in Figure 5.4.

The following are the main features of this component. It is implemented in Java for Android phones and tablets.

- fast access with saved timestamps to reduce delay and facilitate synchronization
- features based on LeGUC features described in [246]
- can help in training machine learning modules for recognition from real MM data
- help answer Research Questions by assuming features are already extracted from MM data

After starting a new session (numbered automatically on each device) and setting a server log-in account, information about the observer like his role (being a researcher, teacher or an educational game developer) and his experience with the observer app are entered. In addition, background information about the player and the session is entered, like the game name and the session ID generated on the other device which has the game running on the capturer app for later synchronization purposes. Some fields for pre-test correct answers numbers are also available if the observer wants to ask the player before starting the actual playing session to compare with post-tests, also administered in the same way.

[1] http://www.firebase.google.com (last accessed in December 2019)

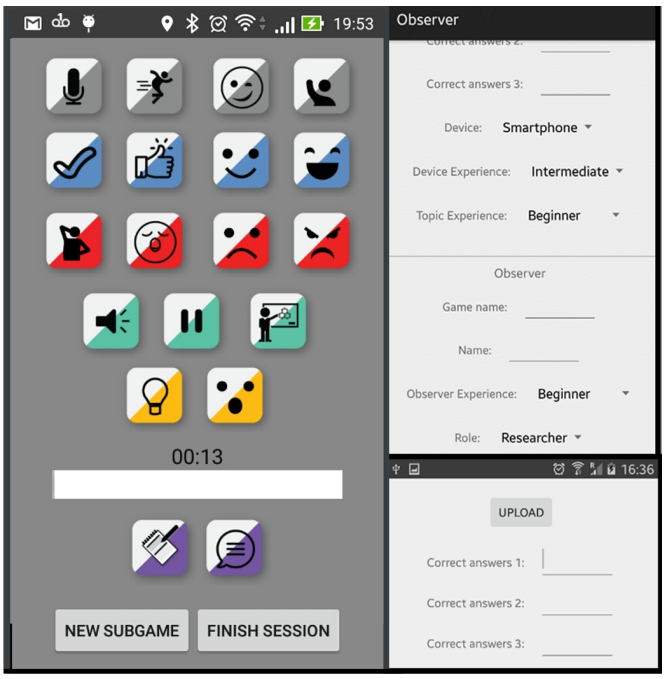



**Figure 5.4:** StoryPlayMM Observer App Screens.

Once the session on the other device has started, the observer starts the session on his device and clicks on the different observations or writes notes which are all saved with timestamps relative to the start of the session. The time lag between starting the game session on one device and the observer session on the other device is later compensated in the desktop evaluation platform. This is realized by providing a slider for the researcher to adjust the offset at the beginning of the timeline until the data is aligned with the observations on the video or the game replay.

After starting the session, a screen is shown with many icons, a textfield and some buttons. The observer clicks on the icons when s/he observes a certain behavior, context or reaction which happens during the testing. For example, the grey icons stand for the tester doing one of those behaviors: speaking, moving, tricking, asking for help. The blue icons stand for his positive reactions like being confident while answering a question, being proud of answering correctly, smiling or laughing. The red icons stand for negative reactions of the tester like being confused, bored, sad or angry. The green icons are for context events like noise, interruption, or you offering help to the tester and explaining something in the game. The yellow icons stand for neutral reactions like reflection and surprise. The observer can use the textfield for two things: taking a note about something (e.g. a bug) and then clicking the left purple button, or writing something that the player said and then clicking the right purple button. The new sub-game button can be used when a player goes to the main menu and choose a new subgame. When the player wishes to finish the session, the observer click on Finish session enters data about post-test if s/he wishes and all the recorded data is saved in an observer log along with its timestamp on the device to be sent to the server.

### 5.2.3 Replayer Component

The Replayer desktop component is the main tool used by the researcher for evaluation. It runs on Windows operating system and its interface is divided into different tabs presenting different data about the sessions. These views are placed in tabs which can be toggled to allow for adjusting the level of detail for each analysis task by expanding the corresponding area. All tabs can be toggled based on researchers needs to avoid overloading the program and the screen when some parts are not needed.

The following are its main features:

- synchronous playback of game events, user video, event log, mobile sensor data, observer logs, profile, session data and model updates (NGLOB narrative, gaming and learning models)
- availability and coherence of replay controls
- coordinated Interactive Navigation of all data based on scenes and events
- summary Statistics, Pre-processing and Visualization
- usable by Non-Programmers
- modular design to accommodate off-the-shelf detectors or researcher-specified rules

The desktop analysis tool is implemented in C# and its interface is built using Windows Presentation Foundation (WPF). Some views of the StoryPlay interface are shown in Figures 5.5-5.8.

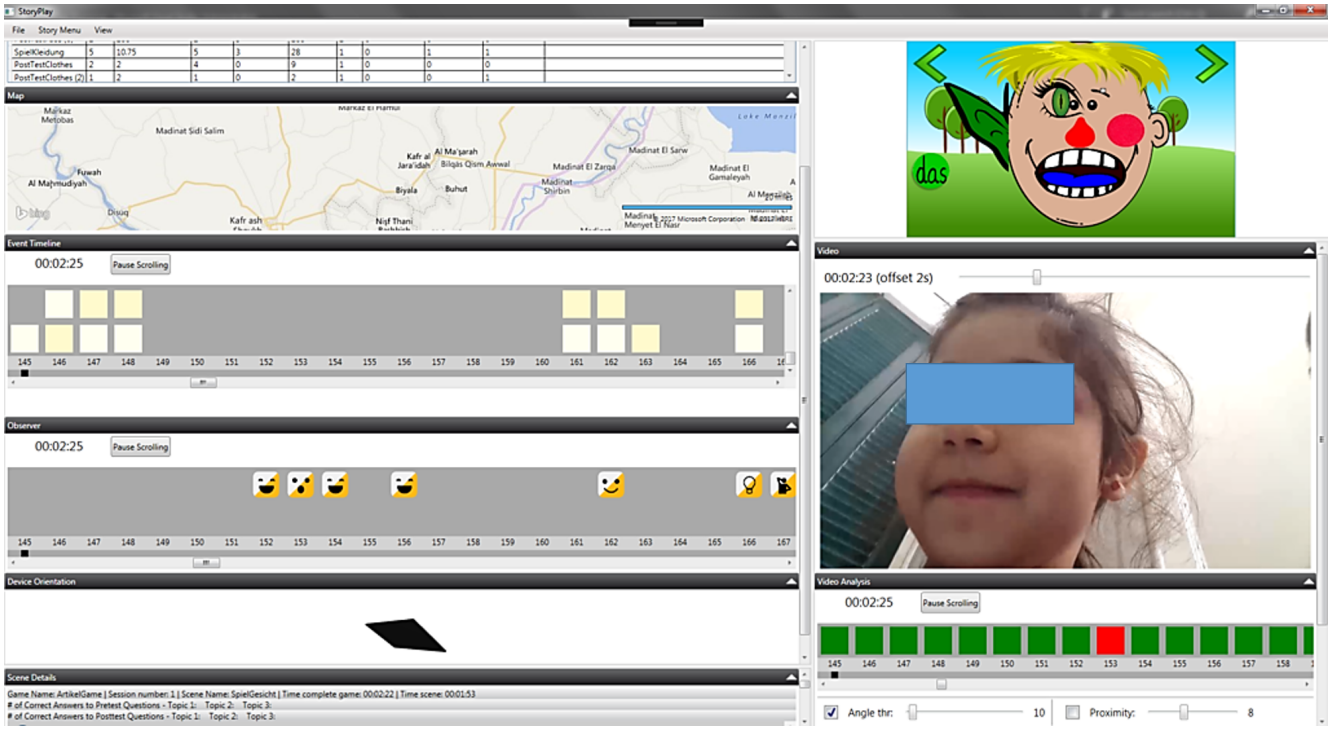



**Figure 5.5:** StoryPlayMM Replayer Component.

### Importing Session Files

After session files were uploaded from the researcher's mobile phone to their own google drive, s/he can log in with his/her account and download session files belonging to him/her. If the app was sent to be tested remotely from participants (after their agreement to send data and choosing which data to send), the data (log files, videos and/or mobile sensing data) is sent to the firebase server to the researcher's account. The observer app files are also sent from the researcher's device and collected on the server. The session ID helps to match observer files with game log files and videos. These files can be imported in the desktop evaluation component on the researcher's computer by clicking sync or by directly downloading and importing the files. Files belonging to the same session are identified by the session ID in the file name and grouped together for investigation.

### Scene Information and Statistics

After the session files are loaded, the different tabs show details about the session. In addition to a tab with general session information like game name, session number, total time on the game, current scene (during replay) and time on the current scene, there is a separate tab with more detailed statistics relevant for research. These are gathered from the log data and displayed in a separate tab with the option to export the data to .xls extension. These data include the game scenes visited in their respective

order, the number of clicks in each scene, the average time per click in each scene, the initial lag before clicks, the missed clicks (clicking on parts of the screen which provide no action), the total time spent on each scene (all visits of the scene accumulated), the number of visits and the number of observations recorded by the observer app on this scene categorized in behavior, context and reaction observations. Examples of a session's gameplay statistics are depicted in Figure 5.6. Examples of how these statistics were used in session evaluation will be discussed in the next section.

File Story Menu View

Scene Statistics

| Scene | Clicks | Average Time/Click | Initial Lag | Missed Clicks | Time Spent | Visits | Behavior count | Context count | Reaction count |
|---|---|---|---|---|---|---|---|---|---|
| Intro | 1 | 22 | 22 | 2 | 22 | 1 | 0 | 0 | 0 |
| Menu | 7 | 2799 | 4 | 5 | 2799 | 8 | 1 | 0 | 0 |
| SpielGesicht | 56 | 100.725490196078 | 5 | 71 | 1658 | 5 | 0 | 5 | 11 |
| PostTestFace | 5 | 2 | 13 | 5 | 448 | 4 | 0 | 1 | 1 |
| PostTestFace (2) | 3 | 1 | 4 | 2 | 288 | 2 | 0 | 0 | 0 |
| PostTestFace (3) | 3 | 3 | 3 | 7 | 291 | 2 | 0 | 0 | 0 |
| PostTestFace (4) | 12 | 3.3 | 2 | 2 | 307 | 2 | 0 | 0 | 1 |
| PostTestFace (5) | 2 | 283 | 0 | 1 | 283 | 2 | 0 | 0 | 0 |
| PostTestFace (6) | 2 | 288 | 2 | 3 | 288 | 2 | 0 | 0 | 0 |
| SpielKleidung | 5 | 10.75 | 5 | 3 | 28 | 1 | 0 | 1 | 1 |
| PostTestClothes | 2 | 2 | 4 | 0 | 9 | 1 | 0 | 0 | 0 |

**Figure 5.6:** Examples of Log Statistics.

### Session Replay Tab

In the Replay area, all interactions are reconstructed from parsed logged data using the StoryPublish serious game engine by parsing the .icml file of the tested game and combining both. This is why the path for the game files has to be chosen at the beginning of the replay. A mouse icon is being displayed at the position of the mouse calculated from mouse movement. This will be extended by changing the color of the mouse icon for different states like mouse clicks. The replay speed can be changed, affecting all other multimodal data replay which runs simultaneously with the in-game events in the different views. One can also choose to jump to a certain event or video frame at any time by clicking on the displayed event icons. As the original log contained all events, the events of interest with meaningful interaction during the game were first identified. The session timestamps where the mouse was just moving around without interaction while playing the game were not considered as meaningful interactions. Thus for skipping a log entry was considered only after a triggered stimuli, i.e. when the user clicked a button to interact with the game.

### Events Timeline Tab

In the events navigator, the user sees a list of significant events moving with time. This is improved by using color codes for different event types and can later also use icons. The events movement is synchronized with the game replay. These events include game generated events like starting a new game or transitioning between scenes, and user events like clicks, triggered stimulus, game pause and resume (when the application is interrupted by the phone, for example.). When hovering over the events on the timeline, the names of the events are displayed underneath. The user can also click on any of the colored squares representing events to jump to this part in all open views simultaneously for a closer investigation.

### Observations Timeline Tab

In this tab icon representations of all observer recordings which are parsed from the observer app log files of the chosen session are displayed on a timeline using the same icons of the observer app which

corresponds to the LeGUC states depicted in Figure 4.8. These move in synchronization with the rest of the representations in open tabs according to the timestamps. When hovering over the "notes" icon, one can see the text written by the observer as a note at a certain instant. Other icons represent any logged reactions like smiling, laughing, showing confusion, boredom, sadness, resentfulness, confidence or pride when answering questions, reflecting or being surprised, behaviors like speaking, moving, tricking or asking for help or context events like noise, interruption or offering explanation by the observer.

#### Session Videos

Video clips are displayed in the WPF GUI using an HTML5 tag called media element tag which supports a broad range of media elements to avoid using extra plugins. Microsoft Expression encoder is used to handle the video files. For synchronizing video with replay, video log events were added to the log-file at the appropriate positions. As the original log contained all events, the events of interest with meaningful interaction during the game were first identified. The session timestamps where the mouse was just moving around without interaction while playing the game were not considered as meaningful interactions. Thus, a log entry was added only after a triggered stimuli, i.e. when the user clicked a button to interact with the game. Initially, a special "EventVideoPlayback" log entry was added for replaying specific chunks of the log. This entry has the following format: DateTime (timestamp to write the video events), EventVideoPlayback, PathToSavedFile (path of the captured video file) Offset (for synchronizing the playback of video while playing a specific portion of the log file) and SpeedRatio (playback speed with 1 for normal playback). As in the original replay, a timer is started for all the events in the log to execute them accordingly. To replay a session part between two important events, all events prior to the event selected in the navigator are executed without timer and then the timer is started from this event to the next significant event. The aforementioned offset tag saved in the EventVideoPlayback entry was initially used to determine from where to start playback and when to stop it, later the synchronization was found to be better when using events like clicks for playback rather than these events.

#### Video Analysis Tab

As discussed earlier, some mobile sensors can give a good indicator when to look at multimodal data like video and when the quality might not be good enough or provide important information about the context of the player. By just using low-power sensors like illumination and gyroscope, bad video frames can be discarded without the need for complex recognition algorithms. The video analysis tab is dedicated to adjusting different thresholds of mobile sensor data. Based on these settings video frames are flagged as good or bad depending on context conditions as can be seen in Figure 5.7. In this figure the testing children put their finger on the smartphone camera and thus covered their faces. Using the illumination sensor with a threshold of 10 these frames were automatically flagged with red squares meaning that they are not usable and can be discarded. However, these frames are not automatically discarded from the beginning as the objective of this tab is to allow researchers to experiment with the best threshold suitable for their particular experiment conditions. A similar method can be used for discarding video frames where the camera is pointing upwards to the ceiling using the angle threshold. For the shaking of the device, a frequency threshold (for shaking speed), an amplitude threshold (for shaking intensity) and a window size can be chosen to flag frames with considerable shaking. The user can freely adjust the default values to his/her own conditions. The color used for good video frames which are more likely usable for evaluation is green. The timeline of these colored squares corresponding to video frames also moves with time in synchronization with the video.

#### Internal Model Changes Tabs

The StoryTec Authoring tool offers game creators the ability to adapt their games by annotating every scene on three dimensions: learner, player and story model (see [82]). Individual learning skills are modeled based on the *Competence based Knowledge Space* [128]. Playing preferences of different players

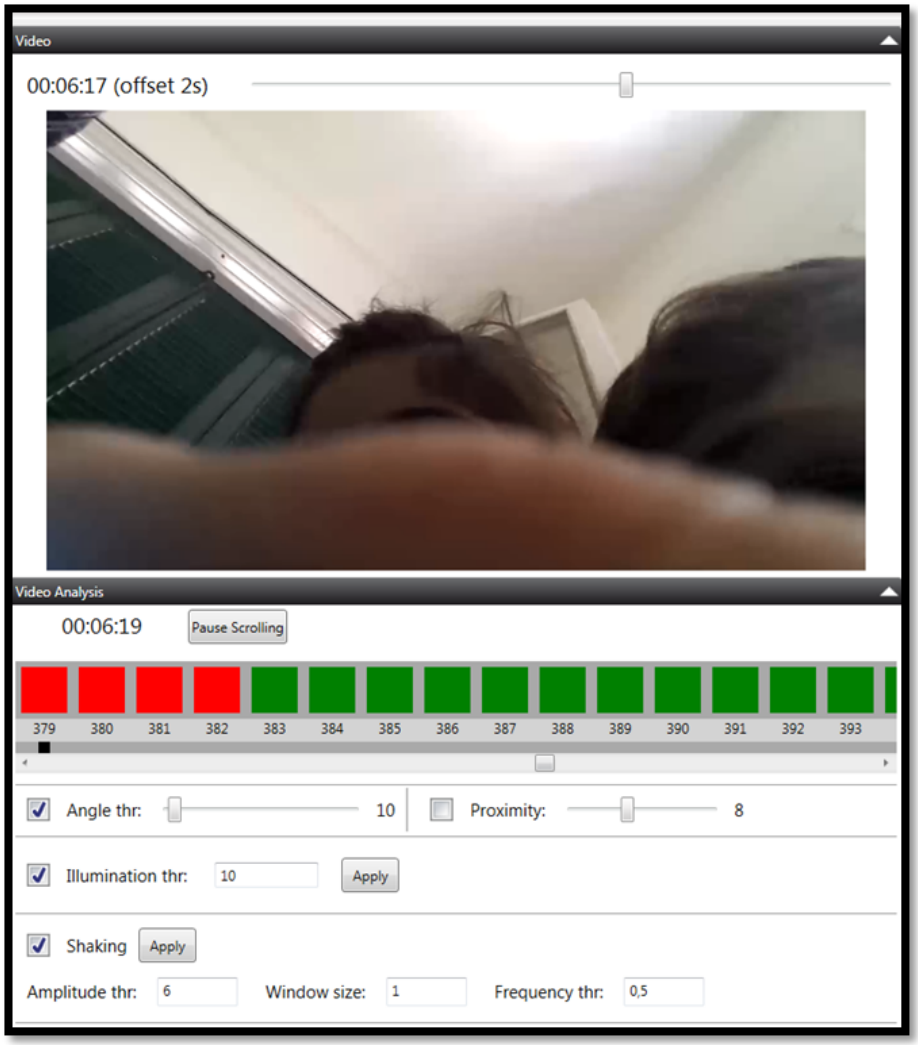



**Figure 5.7:** Flagging Frames with degraded Video Quality based on Mobile Sensing Data.

are modelled based on Bartle's four player models described in [24]: killer, achiever, socializer and explorer by numbers in the interval [0,1] to offer a percentage mapping as players usually show features of different models combined together. The story model is based on the Hero's Journey as modified in [81]. These three models are constantly updated during play based on the game author's annotations made in StoryTec to choose a scene which suits the player best.

In StoryPlay, the state changes of the underlying learner and player model as well as important information about the story path in addition to all current values of active variables are communicated using different visualizations with each event and can be used for various evaluation purposes. This is also possible due to the close coupling with the StoryTec authoring tool which is based on the same internal model (NGLOB) and thus can track data based on it. In Addition the History Tab shows the visited scenes of the game in a graphical representation (see Figure 5.8).

### Mobile Sensors Tabs

A map tab visualizes the map and displays the GPS information captured during the session. Some technical difficulties were faced in using Google Maps so Bing Maps was used. Both required an account to generate a key to be inserted in the code for the application to work. The orientation of the device is represented on a separate tab with a dynamic 3D Model representation of the device. This model moves according to the movement which was recorded in the sensor log entries. Colored arrows show the different directions.

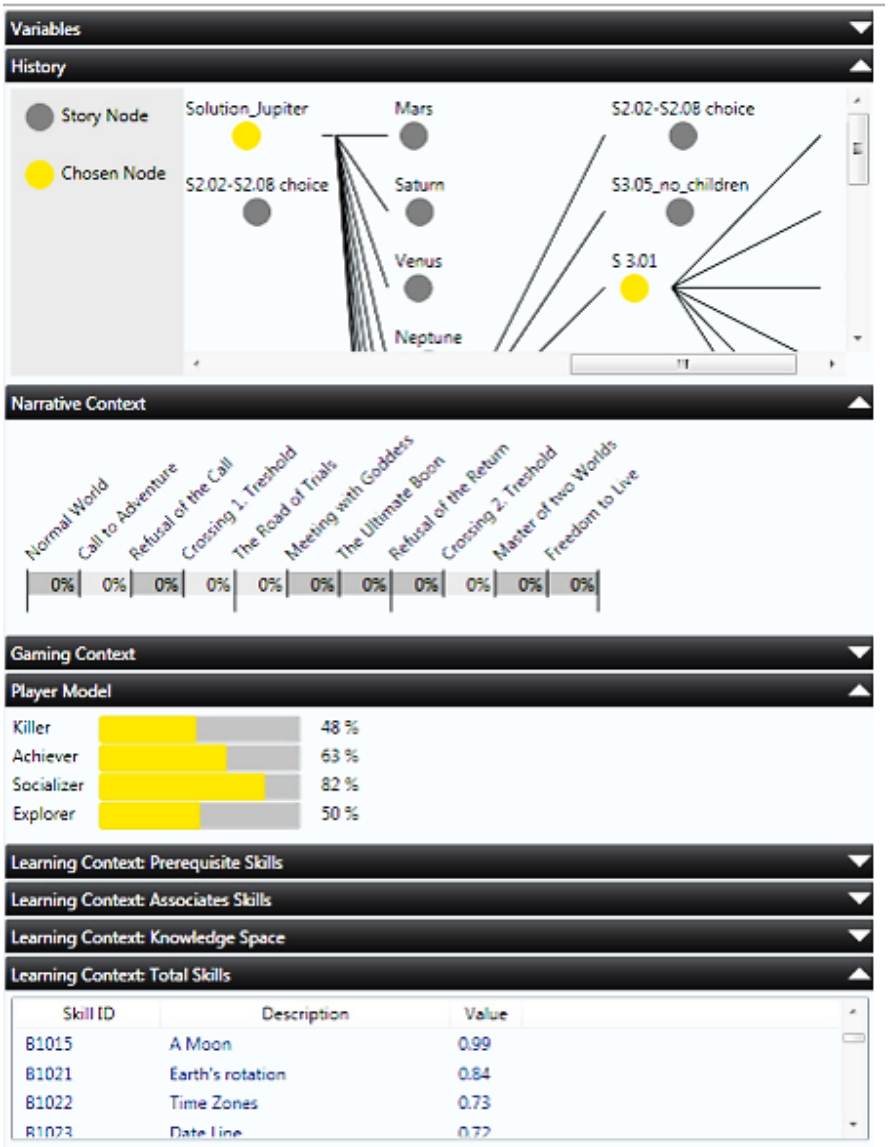



**Figure 5.8:** Scene History, Narrative, Gaming and Learning Context Tabs in StoryPlay [165].

## 5.3 Challenges

### 5.3.1 Synchronization

The advantage in our approach of using built-in mobile sensors is that most of the recordings are done on the same device, so that there are no different hardware sensors requiring some hardware synchronization like in most similar software. However, mutlimodal synchronization was still a major challenge faced in the implementation phase as also reported in literature[38, 66, 75, 240]. Recordings have different frame-rates, start and end time as well as varying reaction time of the observer in case of observer logs. Other synchronization problems were related to StoryPublish which is used to replay the game from the log file. The replay here is an emulation of input and feedback which has its own lags, and the original implementation thus does not allow instantaneous skipping to an event, but rather replays all events quickly until it reaches the desired game state instance when jumping to an event. In addition, rendering the game has its own additional lag and the irregular frame-rate in wpf and decoupling of threads made an exact synchronization almost impossible. Jumping to the correct frame when clicking

on an event was not an easy task, especially with the high frequency of events in a session. Also, cursor position and global time needed to be updated on all tabs when skipping. These calculations were found to be consuming more than 85 percent of the UI thread, making the skipping very slow. Some skip took up to 20 seconds which is why a loading icon was added in this case to avoid confusion. Many optimizations were done and others can still be added to the code to make this process faster. In addition to reducing the number and frequency of such calculation operations and reducing rendering, it can be useful to save the current status in order not to start calculating positions and sizes of all elements in every frame. However, the saving operation might also introduce additional delay.

To address the synchronization issues, an optimal affordable granularity of integration has to be defined for the specific case, then choosing a suitable frame-rate and letting all data obey the same timer. In our case, the timer of the game was chosen in order to introduce the same lags of processing into the other timelines to avoid getting out of sync. To compensate for different starting/ending/reaction times a user-defined offset is allowed using a sliding bar. This can also be improved to automatically align data using some recognition techniques. Event-driven synchronization could also be used to synchronize observer data with the log files.

Game logs and observer logs were also joined on the statistics tab to provide a mapping between quantitative and qualitative data in statistics and not just in replay. One issue in doing this was the time lag between both logs because of the reaction time of the observer. This was also handled in the replay synchronization by adding a manual offset. Here it is required to make sure the lag is not too big to give wrong results. For example, to count number of reaction observations in a certain scene or scene type there has to be a mapping between scenes and the observations made in them which when depending on time only could be inaccurate. Although a "new subgame" button was added in the observer to annotate the beginning of a new part, this information was not reliable as the observer needs to click on the button in time which is not always easy when the game advances fast. So the manual offset set in the replay by the researcher at analysis time need also to be used in the statistics.

### 5.3.2 Interoperability and the Heterogenity of Devices

The heterogenity of devices was also a very time-consuming challenge as applications have to be tested on different devices and different Android versions and updates. This renders it very difficult to make sure it will run smoothly on all possible user devices. Even sensor configurations can differ between devices as some give more useful detailed values than others. The possible values and thresholds for proximity, for instance, differ between devices. The screen resolution problem of devices also affects the replay of games and the calculations of game elements positions and click locations. Some changes may even affect the whole game replay as clicks may be missed when regenerated in replays. Furthermore, video and picture formats supported are different on devices which makes some game elements not run on certain phones. Dealing with permissions to start the camera or save log files was different from one device to another. Thus, it was difficult to give general instructions to all users on how to operate the application and various tests needed to be run to discover these differences. Even after enabling these permissions, some cameras would not work on some devices and the code needed to be debugged for those cases. And even after running the camera, the differences in encoding the data made some videos get saved without the audio, or with audio only without picture. Thus, the heterogenity of devices is considered one of the main challenges for the creation and maintenance of such multimodal applications for mobile devices. Not only this, but compiling the original Storypublish c++/haxe code for Android takes a very long time which makes updating the code with any new feature or adjustment and retesting on devices a considerably time-consuming task.

### 5.3.3 Data Quality

One main challenge of combining multimodal data for the evaluation was concerning the quality of recorded data, for example the quality of video recordings which is also discussed in [143, 94, 169, 131]. Videos from front-facing cameras on smartphones suffer from a dynamic environment which results in variations in illumination, stability, orientation, exposure and distance. To deal with this problem, the aforementioned feature of using data from low-power sensors like accelerometer, orientation, illumination and proximity and applying user-defined thresholds to determine if context conditions are suitable for obtaining good quality data from video or audio was implemented. The goal is to emphasize only useful data to reduce data size and/or reviewing time. Another good feature which could be added in the future would have the goal of enhancing the video using these sensor data, for example automatic illumination compensation in bad video segments. A plugin using ffmpeg, for instance, can be added to provide the user with the option to adjust adjust brightness, contrast and rotation. The recognition of bad orientation could also be done on-the-fly during recording to alert users to adjust the camera view to show their faces in the videos, for example as in Figure 5.9. Face detection could also be used in the beginning of the sessions for this purpose.


**Figure 5.9:** Possible Decision Tree for Predicting Best Media Source.

### 5.3.4 Data Granularity and the Heterogenity of Scenes and Games

One big advantage for this project was the similarity of game structure as all games are created with the same authoring environment StoryTec. Nevertheless, there were still differences between games and scenes. This is why considerable time needed to be invested in making data ready for analysis by the researcher. Pre-processing data so it can give meaningful clues and can help to distinguish significant events while at the same time staying general enough to accommodate different games is not a trivial task. One small example was calculating the time-per-click metric. Although it should be a straightforward task, some differences had to be taken into consideration. Some scenes have some audio instructions at the beginning where mostly the interactions were delayed. To account for this, the initial lag before the first click was not calculated in the average time per click, but as a separate metric. Other scenes contained only a video so there were no meaningful clicks (only missed clicks, i.e. touch down and up without a stimulus invoked). Here the time-per-click metric would be equal to the time-on-scene and would help the researcher only if it is crucial to know if kids skip videos early, for example. In some cases this points to the fact that they find the video boring, or just that they skipped by mistake, which in both cases need to be considered in design. In addition, it can help distinguish the taste of different genders or ages of players for certain content. Also, some scenes can have a transition without a click (e.g. based on a timer) or reach the end of the game, so here the last delta time calculated would be until the end of the scene. Many other similar examples of differences between scenes were encountered in this process.

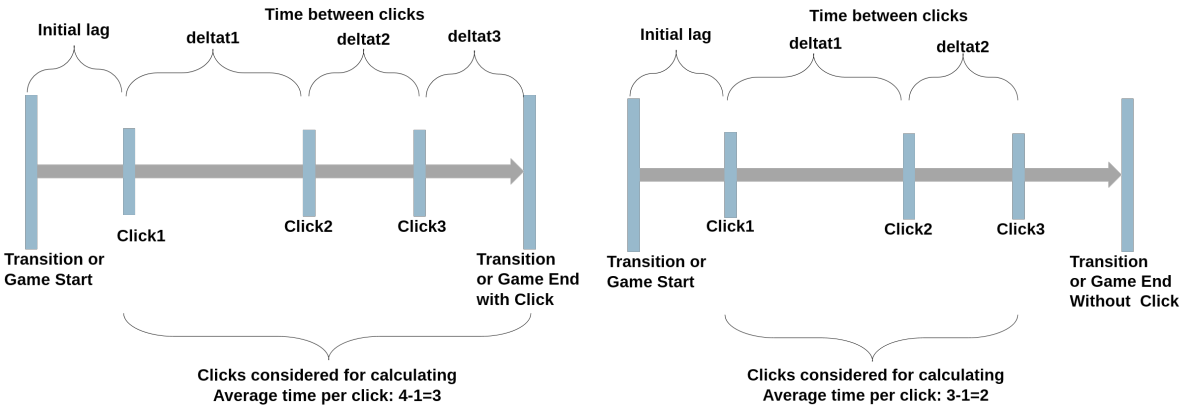



**Figure 5.10:** Differences in Calculating Average Time per Click.

Another example of data granularity was the time-on-scene metric. It is desirable to know the average time spent on each scene (accumulated over all visits of the scene and divided over the number of visits) but also to know the time spent on a particular scene in a particular visit. Also, there is a difference between the time on the first visit and the time on further visits of one and the same scene. This depends on the type of the scene but generally the first time a scene is visited there is more time needed for discovery as in next visits the player has usually already "mastered" this scene. Thus aggregating data sometimes hides important information which is only uncovered when investigated separately. Both options need to be given to the researcher for different purposes.

### 5.3.5 Privacy

The privacy issue was addressed in the initial requirements by giving users full control over sensor activation and over what and when to store and/or upload to the server. However, some additional points concerning privacy needed to be handled. At the beginning, the only way offered for uploading data was uploaded to a firebase project created for gathering log data from this particular application (a key for each app is needed to be added to the project using the account created for the current research). The same process should be applied by any researcher before compiling each of his games so that the data is sent to his/her account which would be the most secure option. Another option which is provided to researchers in the current app is to create their own account inside the created firebase project and to use it for uploading (from capturer) and downloading (by replayer) their data privately without accessing other researchers' data. However, in this case the creators of the initial account (i.e. of this research project) still have access to all data sent to the server. This is why a google drive option was added to the application where users can log into their own google drive accounts and upload data there, so they can have full control and privacy. The problem with both log-in features is the complicated setup needed for them to run on the user end for newly created games as new keys need to be generated for each app and added before compilation. The final more secure option is to use an offline mechanism for transferring the saved files to the researcher's desktop. However, this needs expert users to deal with getting this app data from their devices and using them on their laptops for the replay or sending them to a remote researcher.

### 5.3.6 Other Challenges

In addition to the discussed main challenges, many other challenges were identified and addressed in the initial project requirements, like avoiding obtrusiveness. The problem of high battery consumption of the capturer app was not very concerning as the mobile edugame sessions are relatively short. This

would need to be tested for prolonged sessions where a player repeats the same game several times over a long period of time. Another challenge was making the platform easily usable and intuitive for non-professional users. Many layout changes were carried out to accommodate for the many tabs which may need to be opened at once and to give the user the choice to change the layout to one which suits his/her needs and devices. A switch between horizontal and vertical layout and a flexible tab size in all directions was found to be very useful. Also, making sure the menus, buttons and icons on timelines are descriptive enough to make it easier to understand during analysis and providing easy means of skipping needed many iterations and tests. Other additions like muting videos when their tab is not expanded and muting game sounds while skipping were made to make the evaluation session run more smoothly.

## 5.4 Conclusion and Future Work

In this chapter an environment offering capturing, synchronization, replay, pre-processing and interactive navigation of multimodal data for Serious Games evaluation was introduced. This unified visualization of quantitative and qualitative playlearner data makes it possible to discover relations between game elements and playtester behaviors, affective and cognitive states as well as evaluation context. The steps described for creating this proof-of-concept software, also published in [237, 241, 242] answers parts of Research Questions 3 and 4: How multimodal data can be captured unobtrusively for Serious Games evaluation, how it can be linked to recorded log events and what the associated challenges are. Many improvement possibilities as well as useful applications of the current project have been mentioned throughout this chapter which can be summarized as follows:

- training machine learning modules to predict whether a frame contains faces using the accelerometer and gyroscope sensors
- using built-in smile detectors or facial feature recognition modules
- collecting and saving more user data and demographics automatically and tracking learning progress over time, also the possibility of using information for game rating or recommender systems
- detecting and highlighting other relevant data such as correlations, repetitions and significant events and sequences - detecting event sequences may be helpful for comparing the input stream with some target sequence (e.g. of an expert player) or emphasizing certain pre-defined player behavior. Aggregations and Correlations can be done across users for extensive data for identifying common paths and interaction patterns or rating scenes(e.g. scenes where most users had a sudden movement or smile or laughter etc.) or across modalities for one and the same user for intensive data (when more than one data source has significant change at the same time)
- recorded observer reactions, behaviors and context events can help in training a machine learning algorithm for extracting features from multimodal data by serving as a ground truth
- improvements on the design of the desktop replayer app are mainly in optimizations in replay and synchronization mechanisms. Observer and interaction data can be automatically aligned together instead of the manual offset by training a machine learning algorithm to detects lags in observer session start time. The same can be applied to aligning statistic aggregations.
- video quality can be enhanced using the sensor data monitored, for example automatic illumination compensation in bad video segments - a plugin using ffmpeg, for instance, can be added to provide the user with the option to adjust adjust brightness, contrast and rotation. The recognition of bad orientation could also be done on-the-fly during recording to alert users to adjust the camera view to show their faces in the videos. Face detection could also be used in the beginning of the sessions for this purpose

# 6 Evaluation

In this chapter the evaluation process and results of the implemented platform as well as an application with different serious games and testers are presented. Parts of this chapter are published in [237, 239].

## 6.1 LeGUC Framework Application on Serious Game Evaluation Data

One of the early motivations for the present research on Serious Games Evaluation arose from evaluation studies carried out as part of a research on child-centered design of educational mobile games for preschoolers [243]. A literacy game was implemented using the Pre-MEGa framework, a framework developed in the course of this research for defining quality metrics for fun and effective games developed especially for the age group of preschoolers [245]. When it was time to evaluate the game with real users, preschoolers, a series of evaluation sessions were carried out as described in [243] relying mainly on observation and note-taking (based on a prepared form) which had been found to be the best way to collect evaluation data with this age group [154].

As it is difficult to get accurate verbal data from children for evaluation, observation of their interactions can give valuable feedback for researchers. The notes taken during these evaluation sessions were used for making further game updates improving design and usability of the game and even adding additional sub-games and quizzes. The process of mapping children's reactions to an assessment of game elements back then needed a more systematic method which was not found in literature. This was one major early motivation for the creation of the current theoretical framework with the aim of categorizing each interaction/reaction of Serious Games users into a meaningful category or dimension.

However, the LeGUC model was then developed as a general evaluation framework for Serious Games with special focus on taking into account affect, cognition and context. These three, as shown in the previous sections, need live observation to be recorded, or, if not possible, multimodal capturing of the experience as in StoryPlayMM. Relevant literature was reviewed regardless of the age of Serious Games users collecting and restructuring categories from relevant evaluation frameworks and without taking into account the specific data gathered before as part of the old project.

For an application of the LeGUC theoretical framework on real data, it would not be wise to set up a new evaluation experiment with the framework in mind, as then there might be a bias in the way data is collected to reflect the importance of the chosen categories. Instead, applying this framework on the old data gathered during evaluation sessions can be a good way of bringing this theoretical construct into practice. As attempting to translate the available researcher's notes of children interactions and reactions while evaluating the mobile literacy game revealed interesting in-practice examples which might be helpful for future research, it was decided to include some of them here as a case study. This would serve as a first step for validation.

First, we examine interaction data gathered which could be equally detected using logging without observation or additional multimodal data. One interaction pattern which was a good indicator for an enjoyable sub-game was children repeating this particular sub-game several times before or after proceeding to the next. In our case, this was a sub-game where they could steer a car by tilting the mobile device (using the built-in accelerometer sensor) to catch coming letters. This was attributed to the interaction paradigm used and confirmed the researchers' intuition that it would be more engaging than the more traditional interaction paradigms used in the rest of the sub-games. Thus logged repetition or skipping of a sub-game can be a good indicator of successful design and can thus be translated to game design decisions. This can be regarded as part of the experience category on the gaming dimension on the presented model.

It is interesting to note that even here pure logging would not be sufficient to make sure that children repeated this sub-game because they liked it more and not because they were not able to move on


**Figure 6.1:** Evaluation Sessions at the Nursery.

because of other usability problems (finding the correct button to click on for proceeding, for example). Although this was not the case with our game as it automatically proceeded to the next sub-game unless the user deliberately chose to repeat a sub-game, it might be the case with other games with the lack of additional observational data.

Other data which can fit into this category of loggable actions in the R& R Model were correct/wrong answers given during an Alphabet quiz integrated into the game. This is an example of the performance category on the learning dimension of the framework. An example of a performance action which can reflect a problem on the usability dimension was also found in the observation notes: Some children tried to drag letters during a song which started after a sub-game with dragging letters. They mistakenly repeated the same action because of the similarity in scene design between the game and the song which revealed a hidden interaction design flaw. A similar interaction issue was found when children tried double clicking on an Alphabet range which was designed to cancel selection when re-selected. This would also be inferable from logging if the time between the two clicks is taken into account but would also be more easily recognized with direct observation of interactions/reactions. Another similar issue was found when dragging a letter while holding the device with the other hand touching the screen did not work, which, depending on the logging nature, might or might not be captured.

Moving on to information which cannot be gained through logging alone, the first thing which is found is children's verbal expressions during play (even without asking them to think-aloud, unlike most sessions with adults). Remarks during and after play like: "I did it!", "I won!", "I can't", "How do I move it?" or "Can I play again?" were found to be very valuable for evaluation. According to the utterance, these can be used on each of the three dimensions, respectively, and will need audio recording to be captured in a remote evaluation, or otherwise get completely lost. Even more interesting/unexpected verbal expressions like talking to game characters, singing with songs or explaining to a friend were noted during the sessions and can be captured as useful audio indicators for e.g. affect and immersion. Non-verbal data which can also be captured using audio is laughter which was also encountered during interaction.

An issue related to context which can also be detected using audio was that sometimes children were not able to hear some audio instructions of the game due to occasional noise in the nursery setting. This could help in a design decision like making volume adaptable to surrounding noise level. Non-verbal indicators which would need video capturing include smiles (in our case e.g. when catching a letter, finishing a sub-game or listening to a song), excitement when starting a new phase, embarrassment when missing a letter, concentration when answering a quiz question, reacting to a certain sound effect or being surprised by the fireworks effect. All these indicators of affect and cognition are especially apparent due to the spontaneous nature of most young children and can be missed due to failure in verbalizing them on a post-game survey.

It was also found that for some incidents, logging alone might give very ambiguous or even misleading information: Some kids caught letters in the steering game just by chance and were actually not able

to steer the car. This would be discoverable by either observing behavior indicating this or recording and interpreting accelerometer data with the aim of making sure that achievements were on purpose. Another example was pressing the pause button during the steering game by mistake, which also cannot be directly inferred from logging alone. As can be seen in Figure 6.2, we were able to map all valuable observation notes on dimensions of the framework which shows the capacity of the framework in covering issues encountered in real evaluations. Notes in red need observation or multi-modal data to be captured, light-blue notes are related to context and dark blue notes are interactions recordable by pure logging.

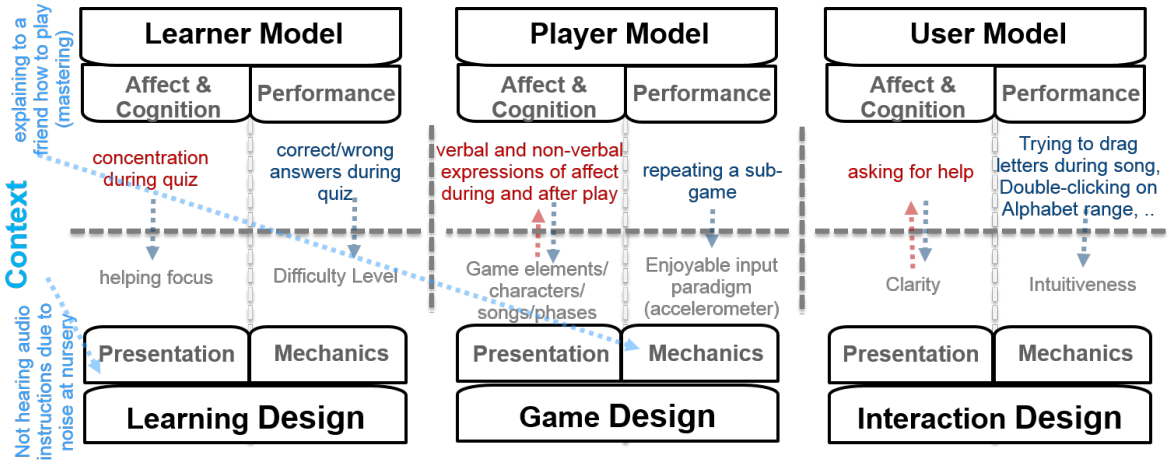



**Figure 6.2:** Mapping Evaluation Data to Framework Dimensions.

## 6.2 Field Study on LeGUC States

In this section the use of LeGUC states described in Figure 4.8 is demonstrated for evaluating learner experience in a mobile educational game for children teaching them the German "Artikel", the Artikel-Game, to give insight for further game development cycles. The game was created with StoryTec [166], an authoring environment for creating scene-based serious games based on the NGLOB (Narrative Game-Based Learning Object) Model [81]. A field study carried out on three days with twenty children in total aged four to ten playing the ArtikelGame on different devices (laptop, tablet, smartphone) in a language learning center was carried out (see Figure 5). The researcher observed and took note of occurrences of different states related to the three categories (Affect and Cognition, Context and Behavior) during the playtesting sessions using the observer mobile app described in Section 5.2.2 designed specifically for this purpose. In this app it was possible to record observations which are saved with timestamps at the moment they are annotated (see Figure 5.4).

The main identified states are depicted in Figure 6.3. The Affective and Cognitive States elicited were: reflection (or concentration, thinking, workload,..), surprise (here positive, e.g. by funny game elements), confidence and confusion (or hesitance), pride and sadness (or disappointment, shame,.., e.g. when giving a wrong answer on a quiz), satisfaction and anger (e.g. anger mostly to usability), enjoyment (or excitement, engaged concentration) and boredom. Thus, our findings on this dimension closely resemble the affective and cognitive states identified in [91, 52], where the authors also identified the occurrence of these states using observation of participants playing an educational game.

To better understand how these states can be linked to event logs recorded from the game, they were further categorized according to the time they took place related to in-game events on the three dimensions of learning, gaming and using. The results are depicted in Figure 6.4.

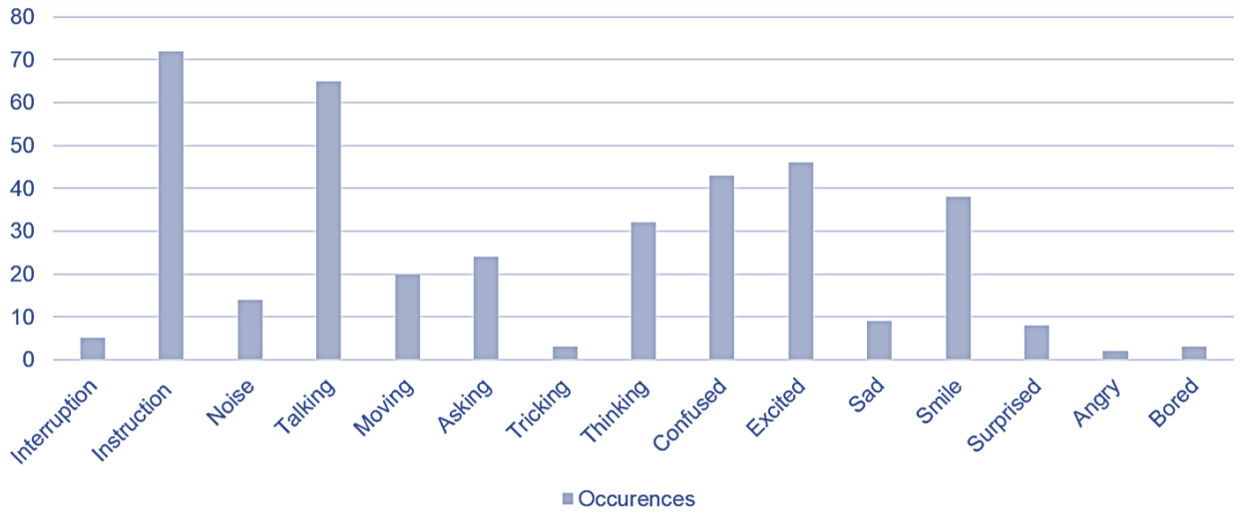



**Figure 6.3:** Occurences of Congnitive, Behavioral and Context States during Playtesting Sessions.

| | State | Learning | Gaming | Using |
|---|---|---|---|---|
| Before Action | | Thinking about answer | Thinking about strategy | |
| | | Understands/confused by question | Understands/confused by mechanics | Understands/confused by presentation |
| During Action | | Confident/hesitant in answer | Confident/hesitant in move | Confident/hesitant in interaction choice |
| After Action | | Proud/sad about learning achievment | Proud/sad about game achievement | |
| | | Helpful feedback? | Emotions elicited by game story | System responds as expected? |
| | | Surprised by feedback element | Surprised by game element | |
| Anytime | | Enjoying/ Bored | | |

**Figure 6.4:** Timing of Affective and Cognitive States Related to Dimensions of In-Game Actions.

Categorizing affective and cognitive states according to their occurrence in relation to achievements is not a new concept and has been investigated in research, resulting in notions such as activity emotions, prospective and retrospective outcome emotions [201, 203, 204]. For the current framework, they are more explicitly defined in relation to Serious Games evaluation and describing how their meaning will differ according to the dimension of the activity.

As for context states and events shown in Figure 6.3, the identified parameters related to environment were location (also indoors or outdoors), illumination (can influence ability to see game graphics clearly) and noise (can influence ability to hear game instructions). Those related to the device were shaking (may make it difficult to interact, e.g. putting the device on the lap and trying to hit an object) and orientation (landscape/portrait, flat/upright influence viewing angle and size). Other external events influencing the interaction are interruptions (e.g. a teacher opening the door and getting into the room) and instruction (the observer helped the student proceed in the game after s/he were stuck and asked for help, to prevent the testing session from stopping early or lasting longer than intended..). Most of the context states were also identified in mobile learning research as presented in Table 3.2 but were lacking or not precisely defined in Serious Games evaluation frameworks and studies found in literature.

In addition to affect and context, other behaviors expressed by the participants were observed which cannot be categorized under any of the two notions (see Figure 6.3). Some kids were speaking with the observer, speaking with friends, answering in-game questions verbally, singing (along with a song in the game or alone), laughing or thinking aloud. All this can be captured in a testing session using a microphone and are important for evaluation. These were grouped under the behavior: speech. Another behavior exhibited was movement: People (and especially children) tend to be moving a lot while engaged in activities and this was observed during the playtesting sessions. Sudden movements affect the device condition and also media captured during sessions such as recorded session videos. They can also lead to undesired game actions like accidentally touching the screen, especially when navigation buttons are placed near screen edges, which is considered a poor design decision for mobile apps.

The behavior of requesting help from the facilitator was considered separately from the speech behavior as it can be used as an indicator of confusion or uncertainty but where the problem faced is explicitly verbalized, offering enormous help to evaluation. Taking note of this can help, for example, in highlighting this portion of speech, if recorded, as important to consider. Another behavior, which is actually related to interactions inside the game, but can be better recognized by observing reactions of a participant is the behavior of tricking or gaming the system. This occurred so frequently in our sessions that it deserved to be considered in the model: A participant is quickly clicking on random answers of a question instead of thinking about the answer until the correct answer is clicked. If during looking at the log data, the researcher only investigates where the user clicked and not how fast s/he clicked, s/he might miss to interpret this behavior and falsely assume a purposeful decision in choosing the answer. Thus recording this behavior can help explain strategies behind pure user event logs. Icons similar to those from Figure 4.8 were then used as real-time annotating buttons in the new version of the observer app to be used for further playtesting sessions of Serious Games. Figure 6.3 shows occurrences of some behavioral, affective and context states recorded during the playtesting sessions.

## 6.3 Evaluation of StoryPlay Multimodal Platform

### 6.3.1 Playtesting with Children by the Researcher

After an initial technical evaluation where limitations were identified across multiple devices and data was collected using them, StoryPlay Multimodal was used to evaluate a game teaching children the German Artikel. To evaluate the application of the tool a special educational game was created with special requirements. The following were its main requirements (see Figure 6.5):

- created with StoryTec

- running on different devices
- diversity in mechanics and media
- collecting high number of different gameplay-related events in a relatively short time
- analytics in mind: using tags for recognizable and quantifiable log events
- evoking affect reactions: funny and motivating elements
- a minimum of three different subgames for comparison
- integrating pre-& post-Test

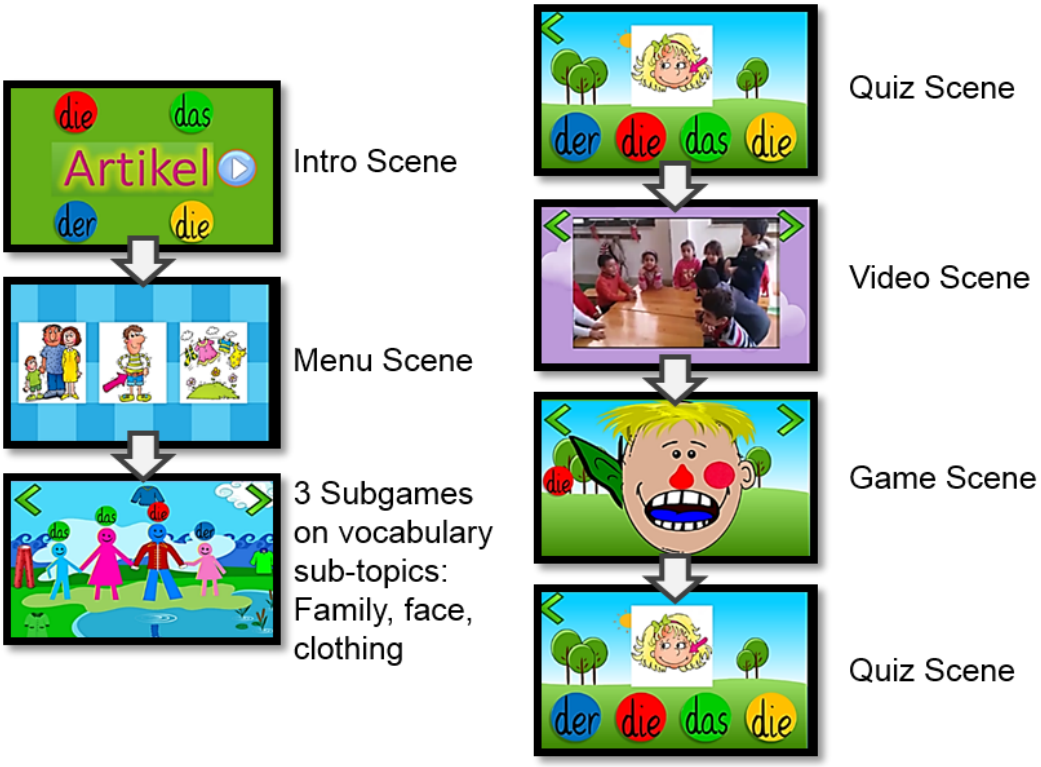



**Figure 6.5:** Some Screens of the Mobile Game created with StoryTec and tested with StoryPlayMM.

The resulting game had three scene types for each subgame (game scenes, quiz scenes and video scene).

Three playtesting sessions were carried out in addition to a remote session which kids carried out with their parents at home. The data was automatically sent to a server from their mobile devices (if they chose to allow this). In total there were twenty unique children, aged from 4 to 10, but many kids played more than one session of the game. The data from the first session was used to improve the game for the second session, simulating iterative design and evaluation of games using information provided from the data. In addition, some initial sessions were carried out without the observer app relying on traditional paper-based note-taking of observations by the researcher to assess the difference in evaluation experience.

Different devices were used in the evaluation a Laptop, a tablet and a smartphone, also different devices were used for the observer app: a tablet and a smartphone - in addition to the participants devices at home. In total, 22 log files were recorded.(See Table 6.2)

A limitation of this first evaluation study is that it was a single case study with one Serious Game created with StoryTec. However, the varying types of scenes and subgames in the game offered variation and depth to the analysis. In addition, the game had to stay fixed to allow for trying the different features on the same game, on multiple user sessions to have comparable results. Also, the observation approach is better to be used with testers on a relatively small scale, then an algorithm can be defined based on the data for larger scale evaluations.

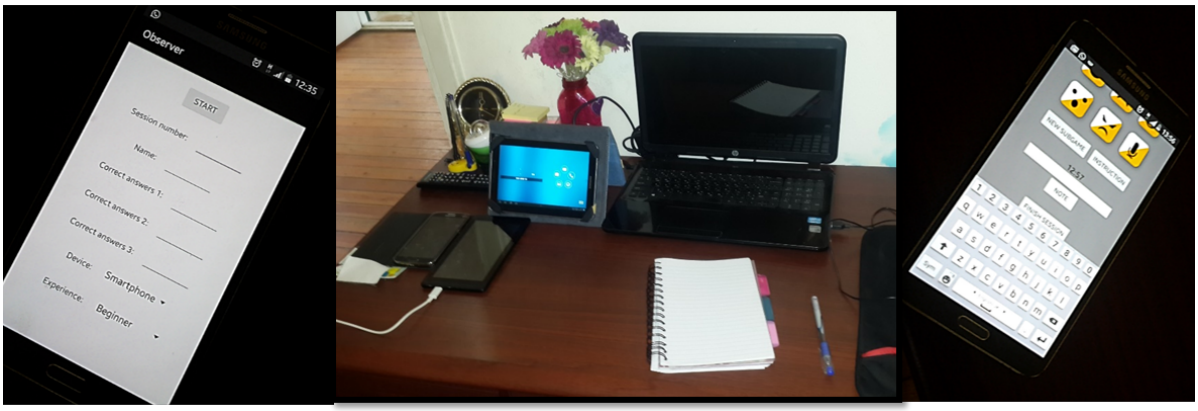



**Figure 6.6:** Evaluation Session Setup.

**Table 6.1:** Evaluation Participants in Different Sessions.

| Session | 4y | 5y | 6y | 7y | 8y | 9y | 10y | Total |
|---|---|---|---|---|---|---|---|---|
| no.1 | 0 | 1 | 5 | 3 | 1 | 1 | 1 | 12 |
| no.2 | 1 | 0 | 6 | 0 | 1 | 0 | 0 | 8 |
| no.3 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 4 |
| Remote | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 |
| Total | 2 | 4 | 8 | 3 | 1 | 1 | 1 | 26 (20 unique) |

**Table 6.2:** Devices and Software Used in Evaluation and Number of Log Files Generated.

| Session | Device | Game Version | Game Log Files | Observer Log Files | Device Videos | Observer Videos | Observer Notes |
|---|---|---|---|---|---|---|---|
| no.1 | Laptop | 1 | 0 | 2 | 2 | 8 | 12 |
| no.2 | Tablet | 1 | 8 | 8 | 8 | 1 | 8 |
| no.3 | Smartphone | 2 | 4 | 4 | 4 | 2 | 4 |
| Remote | Tabl./Sm. | 2 | 10 | 0 | 4 | 0 | 0 |
| Total | | | 22 (14 unique) | 14 | 18 | 11 | 24 |

The evaluation with children helped mainly identify some improvement aspects which were reported in [246] and improved in further development versions of all modules. In addition, the evaluation process was useful for improving the game itself in several iterations.

As a main benefit of the platform was enabling the linking of quantitative with qualitative data, the evaluation platform was used for exploration in this regard. The two main log statistics used for deeper investigation for each scene visited were the number of stimulus in a scene and the time between stimulus. These were examined in relation to the observed states from our LeGUC states. When interesting behaviors were found in the extracted statistics, corresponding observations in the observer file were investigated to see if they give additional information. Some initial exploration showed that the tool can be very helpful in pointing out relations between recorded observations and action logs. For example, it was helpful to highlight differences between reactions to scene types. It was found that depending on scene types and context, more clicks in a scene was not always better and more time before carrying on an action was not always bad. This was in-line with our hypothesis that observing reactions, behaviors and context can help disambiguate game logs.

As our game had different scene types (game, video, quiz), the number of clicks in a scene and the time between clicks had different meanings in different scene types: for a quiz scene it is better to finish the scene quickly whereas in the video scene this would mean that the kid was bored and wanted to skip the video. A high number of clicks in the game scene is a good sign meaning there is a lot of engaged interaction whereas in the quiz scene this means many wrong answer attempts. The number of states recorded using the observer app were measured in different scenes to see if they correlate with the logging events differently. Indeed, the number of clicks and time spent in the game scenes both lead to a higher number of recorded events. Whereas in the video and quiz scenes they had a different influence. For example in the video scene, when more time is spent, less behavior is observed. Observations were necessary to make sense of the logging data. This can be an indicator that multimodal and mobile sensing data will help disambiguate some logging data [196] and help determine context of the experience.

| Correlation Matrix (n=699) | | | | | | |
|---|---|---|---|---|---|---|
| Variable | ObsCount | BehaviorCount | Clicks | ContextCount | gameScene | Initial_lag |
| ObsCount | 1.000 | | | | | |
| BehaviorCount | **0.656** | 1.000 | | | | |
| Clicks | **0.506** | 0.137 | 1.000 | | | |
| ContextCount | **0.713** | 0.269 | 0.353 | 1.000 | | |
| gameScene | 0.360 | 0.203 | **0.624** | 0.242 | 1.000 | |
| Initial_lag | 0.349 | 0.421 | -0.019 | 0.230 | 0.040 | 1.000 |
| Missed_clicks | **0.551** | 0.187 | **0.720** | 0.331 | 0.435 | 0.087 |
| NegObs | **0.623** | 0.273 | 0.255 | **0.648** | 0.186 | 0.204 |
| PosObs | **0.780** | 0.291 | **0.548** | 0.321 | 0.327 | 0.168 |
| quizScene | -0.262 | -0.225 | -0.230 | -0.163 | -0.494 | -0.322 |
| ReactionCount | **0.871** | 0.333 | **0.565** | 0.457 | 0.343 | 0.197 |
| Time_Spent | **0.692** | 0.367 | **0.637** | **0.575** | 0.489 | **0.562** |
| videoScene | 0.114 | 0.132 | -0.063 | 0.023 | -0.049 | **0.510** |

**Figure 6.7:** Extracting and Aligning Count Measures for Each Scene from Observer and Game Logs.

Different count measures were extracted from game logs and observer logs, aligned together and aggregated to count data frequencies of different reactions, behavior and context in each game scene. Count measures extracted for each scene from the observer logs included reaction, context and behavior recordings count as well as recorded observer notes and participant utterances. From the game log, measures such as the number of visits for each scene, the average number of clicks per visit, the average number of missed clicks per visit, the average time per click, the initial delay (time to first click on first visit) and the time on scene were extracted.

To get a measure which can be used irrespective of the scene type, the significance metric was normalized using the respective values found for each scene type.

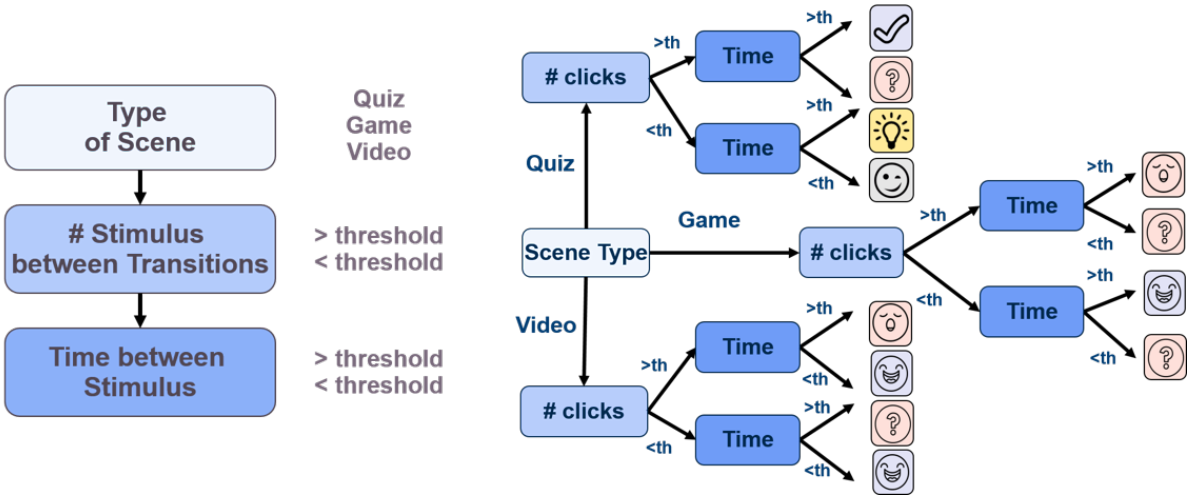



**Figure 6.8:** Detecting Interesting Behavior in Logs and Investigating Corresponding Observer Data.

| | | Quiz Scene | | Video Scene | | Game Scene | |
|---|---|---|---|---|---|---|---|
| **#Clicks** | compared with average | > av. | < av. | > av. | < av. | > av. | < av. |
| **n** | | 361 | 369 | 6 | 7 | 7 | 17 |
| | states | | | | | | |
| **Pearson Correlation** | *coeff* | 0.234 | 0.271 | 0.901 | 0.847 | 0.568 | 0.399 |
| | *p-value* | 0.05 | 0.00003 | 0.098 | 0.016 | 0.184 | 0.11 |
| **Time per click** | states | | | | | | |
| | *coeff* | -0.07 | -0.224 | -0.740 | -0.567 | | |
| | *p-value* | 0.56 | 0.0006 | 0.092 | 0.185 | | |

**Figure 6.9:** Different Relations between Game and Observer Count Measures for Different Scene Types.

Correlation Matrix (n=24)

| Variable | ObsCount |
|---|---|
| ObsCount | 1.000 |
| CPVSignificance | 0.099 |
| TPVSignificance | **0.701** |

**Game Scenes**

Correlation Matrix (n=12)

| Variable | ObsCount |
|---|---|
| ObsCount | 1.000 |
| CPVSignificance | -0.359 |
| TPVSignificance | -0.389 |

**Video Scenes**

Correlation Matrix (n=306)

| Variable | ObsCount |
|---|---|
| ObsCount | 1.000 |
| CPVSignificance | -0.010 |
| TPVSignificance | 0.358 |

**Quiz Scenes**

**Figure 6.10:** Metric Significance (Distance from Average) of ClicksPerVisit (CPV) and TimePerVisit (TPV) Calculated for each Scene Type Separately (to Compare Scenes) and Correlated with Number of Observations.

Some results suggest a relation between these calculated measures and the observation counts irrespective of the scene type as can be seen in Figure 6.11.

Coefficient Estimates: Model 3 for ObsCount (4 variables, n=342)

| Variable | Coefficient | Std.Err. | t-Stat. | P-value | Lower95% | Upper95% | Std. Dev. | Std. Coeff. |
|---|---|---|---|---|---|---|---|---|
| Constant | -0.033 | 0.267 | -0.123 | 0.902 | -0.557 | 0.492 | | |
| ClicksPerVisit | 0.554 | 0.059 | 9.370 | 0.000 | 0.438 | 0.671 | 2.070 | 0.453 |
| NCPVS | -0.925 | 0.249 | -3.717 | 0.000 | -1.414 | -0.435 | 0.493 | -0.180 |
| NTPVS | 0.850 | 1.065 | 0.798 | 0.425 | -1.245 | 2.946 | 0.224 | 0.075 |
| TimePerVisit | 0.003713 | 0.001520 | 2.442 | 0.015 | 0.000722 | 0.006703 | 159.573 | 0.234 |

**Figure 6.11:** Normalized Factors for ClicksPerVisit and TimePerVisit Significance Metric Correlated with ObservationCount.

### 6.3.2 Evaluation by Students

This evaluation study included undergraduate Computer Science students taking a Game Design course in their last semester and was carried out at the end of the course as practical application on the topic of digital game prototyping. Two different classes took part in the sessions. From the first group, seven students actively took part, 4 female and 3 male students, and from the second group 9 students were involved, 5 male and 4 female students. In some milestones, the students worked in groups, in others individually. The evaluation process consisted of the following steps:

1. creating a game with Storytec and exporting it into an apk
2. testing apks of friends on their mobile devices in the lab while another team member acts as the observer using the observer app on his/her device
3. filling out an online form about their playtesting process (see Appendix B)
4. letting another tester (preferably a younger family member at home) test their apk outside the lab while s/he takes notes using the observer app
5. filling out a form about using the capturer and observer app (see Appendix B)
6. the instructor checks the data sent to the server from the devices and assists in technical problems

The main aim of the study was to test the mobile applications on different devices and by different playtesters and observers as well as test the smoothness of the whole process. An added value of this study to the first one is also to get remote observer files and test the procedure of their integration with remote playtesting log files from another device. The online form steps guided students through the evaluation process so that minimum interventions were needed and asked them questions useful for assessing the evaluation experience. The last step of testing the desktop replay component by students is still to be carried out in a further milestone but it was used by the instructor to test the whole integration and view the received files.

Appendix B shows some results from the online forms filled by the students in their evaluation study. It can be seen that one major problem was found with respect to switching on the camera on the devices which make the app crash on some devices as discussed in the implementation challenges in Section 5.3. Although this problem had been resolved on other devices during the initial technical evaluation, it was found that it was still present on other devices. In addition, some devices failed to send the data to the server because of restrictions imposed on apps not downloaded from the Google Play app store. Also, all permissions for the app like camera, location and data access had to be set manually by going to the settings (these steps were explained in detail to the students in the online form but some students failed to carry them out correctly). In general, the challenge of getting all features to work properly on all different devices was found to be a very difficult and time-consuming process and new android updates need further updates in some apps when they emerge.

The majority (80%) of students found the apps useful in testing their friends' games as well as improving their own game by observing others playtesting it. They also found most controls intuitive. A features which was found to need improvement was that clicking the back button while uploading cancelled the

**Figure 6.12:** Participants in Student Evaluation.

uploading while only the home button allowed for continuing the upload in the background. In addition, in both evaluation procedures, the multimodal capturer app helped enable remote evaluation in more naturalistic and unobtrusive settings. Mobile sensing data helped in identifying bad frames from video recordings and give richer information about the session. Using the observer app for the observations also revealed valuable insights into Learner Experience aspects as demonstrated in [246]. According to occurrences of some behavioral, affective and context states recorded during the playtesting sessions, improvements for the next iteration of the observer app were identified and implemented (see more details in [246]).

### 6.3.3 Evaluation by a Teacher

The process of using the three components of StoryPlayMM to evaluate a mobile educational game (Artikelgame) was carried out by a German teacher giving a German language workshop for Egyptian kids in a learning center in Egypt. The sessions included 7 boys and 1 girl between the age of 4 and 7. The following steps were included in the evaluation:

1. Running the game with the capturer app on one device and filling a profile at the beginning of every session for each child
2. Running the observer app on another device and filling a profile about the teacher as the observer
3. Letting each child test the game with the capturer app in the centre and making observations using the observer app while the app is recording multimodal data
4. Opening session data on StoryPlayMM desktop component after the sessions and examining them to give feedback on encountered problems and possible improvements in usability
5. Filling out an online form about using the capturer and observer app (see Appendix B)
6. Filling out an online form about the replayer software

The feedback given by the teacher in the evaluation forms suggests a smooth evaluation session and an intuitive use of the observer app by the teacher. However, it was noted that the sessions were not properly terminated with the StoryPublish app which resulted in many session logs recorded into one file with many pauses in-between sessions, instead of creating a new log file for each session. This means the process of terminating the game between sessions was not intuitive for the teacher.

## 6.4 Conclusion and Limitations

The theoretical as well as implementation frameworks developed and presented in this thesis were put into practice in this chapter by undertaking several user studies. These were carried out to demonstrate

the benefits of combining multimodal data with event logging for the evaluation of mobile learning games. Structuring and linking raw multimodal data for easier navigation was found to be very helpful in carrying out user studies of this type of software. Unified visualization of quantitative and qualitative playlearner data made it possible to discover relations between game elements and playtester behaviors, affective and cognitive states as well as evaluation context. Results helped providing the missing answers for Research Question 4 concerning the benefits of multimodal data for interpreting log events.

However, the same challenges discussed in Section 5.3 were the main challenges faced during the evaluation process. Once these challenges have been overcome, a more thorough evaluation investigation can follow. As discussed in 5.4, it is planned to add some more analysis and visualization features on StoryPlay in the future and continue the current evaluation process to make improvements on usability. An evaluation study with more experienced Serious Games researchers other than the authors of the paper would also be desirable. More research is needed to determine when exactly adding more data adds value for the evaluation of Serious Games and when it is a waste of resources. Some studies in mobile sensing address these problems to adaptively switch on different sensors according to environmental, device and user conditions. This research area is especially promising as it helps develop efficient and effective multimodal mechanisms which add richness to evaluation processes.

# 7 Conclusion and Future Work

The interdisciplinary field of this thesis topic is emerging and can benefit from studies in adjacent fields to accelerate faster. Mobile sensor technology can help solve a problem which always existed in assessment which is the difficulty to gather naturalistic data allowing to track not only users interactions with the software but also the context in which these interactions take place. This can help developers update their software according to new trends and circumstances and help teachers track their students' natural interactions.

After investigating several research fields associated with the research topic, a review of available theoretical frameworks as well as research and commercial tools showed a need for defining theoretical frameworks especially tailoring the use of multimodal data for the evaluation of mobile serious games.

The LeGUC framework for linking serious games log data with affective and cognitive states, user behavior and context data was proposed based on a literature review as well as different studies carried out in the field for evaluating mobile serious games with children and students by serious games researchers and teachers. It has the aim of filling a gap in theoretical foundations for Serious Games evaluation by focusing on the value of adding multimodal data to event logs.

First, dimensions of evaluation aspects were defined based on a literature review. Then it was examined which role multimodal data can play in measuring these aspects, specifically in determining reasons behind users' logged gameplay actions and their responses to game events using the proposed model. Finally, practical examples of using this model for combining logging with multimodal data in evaluation were discussed as well as applied on a case study as a first validation method.

The framework is an attempt to answer parts of Research Questions 1-3 as it defines parameters for a serious games evaluation (Research Question 1), defines a hypothesis of why and when multimodal data is needed to interpret log events in Serious Games Evaluation (Research Question 2) and defines different types of multimodal data needed in evaluations and when they can be used (Research Question 3). It was used as a foundation for developing a multimodal Serious Game evaluation platform as a proof-of-concept in order to show the impact of using (mobile) multimodal data on Serious Games evaluation.

The latter environment offers capturing, synchronization, replay, pre-processing and interactive navigation of multimodal data for Serious Games evaluation. This facilitates discovering relations between game elements and playtester behaviors, affective and cognitive states as well as evaluation context. The steps described for creating this software answers parts of Research Questions 3 and 4: How multimodal data can be captured unobtrusively for Serious Games evaluation, how it can be linked to recorded log events and what the associated challenges are.

Finally, several user studies were carried out to demonstrate the benefits of combining multimodal data with event logging for the evaluation of mobile learning games. Structuring and linking raw multimodal data for easier navigation was found to be very helpful in carrying out user studies of this type of software. Unified visualization of quantitative and qualitative playlearner data made it possible to discover relations between game elements and playtester behaviors, affective and cognitive states as well as evaluation context. Results helped providing the missing answers for Research Question 4 concerning the benefits of multimodal data for interpreting log events.

However, many challenges were faced in implementation and evaluation phase which were explained in detail in the course of this work along with discussing some implemented and some possible future solutions. Further improvements which can be made to the framework were also presented. Once these problems are fully solved, a more thorough investigation can be carried out to validate all research claims with heterogeneous games and users and getting more accurate relations. Hopefully, the results of this research can help start a series of deeper and wider investigations facilitated by the platform provided and extended, for example, using machine learning.

## References


[1] Sam Adkins. *Serious games industry in boom phase: The 2019-2024 worldwide game-based learning market*. Ambient Insight (Metaari), August 2019.

[2] Kingsley Oryina Akputu, Kah Phooi Seng, and Yun Li Lee. Facial emotion recognition for intelligent tutoring environment. In *Proceedings of the 2nd International Conference on Machine Learning and Computer Science (IMLCS)*, pages 9–13, Kuala Lumpur, Malaysia, August 2013.

[3] Naif R. Aljohani and Hugh C. Davis. Learning analytics in mobile and ubiquitous learning environments. In *Proceedings of the 11th World Conference on Mobile and Contextual Learning (mLearn)*, Finland, October 2012. CEUR-WS.

[4] Naif R. Aljohani and Hugh C. Davis. Significance of learning analytics in enhancing the mobile and pervasive learning environments. In *Proceedings of the 6th International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST)*, pages 70–74, Paris, France, September 2012. IEEE.

[5] Anissa All, Elena Patricia Nunez Castellar, and Jan Van Looy. Measuring effectiveness in digital game-based learning: A methodological review. *International Journal of Serious Games (IJSG)*, 1 (2):3–20, June 2014.

[6] Asmaa Alsumait and Asma Al-Osaimi. Usability heuristics evaluation for child e-learning applications. In *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services (iiWAS)*, pages 425–430, Kuala Lumpur, Malaysia, December 2009. Association for Computing Machinery.

[7] Luigi Anolli, Fabrizia Mantovani, Linda Confalonieri, Antonio Ascolese, and L. Peveri. Emotions in serious games: From experience to assessment. *International Journal of Emerging Technologies in Learning (iJET)*, 5(SI3):7–16, November 2010.

[8] Aitor Apaolaza, Simon Harper, and Caroline Jay. Understanding users in the wild. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility (W4A)*, 13, pages 1–4, Rio de Janeiro, Brazil, May 2013. Association for Computing Machinery.

[9] Kimberly E. Arnold and Matthew D. Pistilli. Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK)*, pages 267–270, Vancouver, British Columbia, Canada, April 2012. Association for Computing Machinery.

[10] Ivon Arroyo, David G. Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. Emotion sensors go to school. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education (AIED)*, pages 17–24, NLD, July 2009. IOS Press.

[11] Ivon Arroyo, Beverly Park Woolf, Winslow Burelson, Kasia Muldner, Dovan Rai, and Minghui Tai. A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education (IJAIED)*, 24(4):387–426, December 2014.

[12] Nikolaos Avouris, Vassilis Komis, Meletis Margaritis, and Georgios Fiotakis. An environment for studying collaborative learning activities. *Educational Technology & Society*, 7(2):34–41, April 2004.

[13] Nikolaos Avouris, Vassilis Komis, Georgios Fiotakis, Meletis Margaritis, and Eleni Voyiatzaki. Logging of fingertip actions is not enough for analysis of learning activities. In *Proceedings of Workshop on Usage Analysis in Learning Systems in the 2005 Conference on Artificial Intelligence in Education (AIED)*, 7, pages 1–8, Brighton, UK, 2005.

[14] Kiavash Bahreini, Rob Nadolski, and Wim Westera. Multimodal emotion recognition for assessment of learning in a game-based communication skills training. In *Proceedings of the Workshop Learning Analytics for and in Serious Games at the European Conference on Technology Enhanced Learning (ECTEL)*, pages 22–25, Graz, Austria, 2014.

[15] Kiavash Bahreini, Rob Nadolski, and Wim Westera. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments*, 24(3):590–605, April 2016.

[16] Kiavash Bahreini, Rob Nadolski, and Wim Westera. Towards real-time speech emotion recognition for affective e-learning. *Education and Information Technologies*, 21(5):1367–1386, September 2016.

[17] Ryan S. J. D. Baker and Kalina Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining (JEDM)*, 1(1):3–17, October 2009.

[18] Ryan S. J. d. Baker, Sidney K. D'Mello, Ma. Mercedes T. Rodrigo, and Arthur C. Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, April 2010.

[19] Ryan Shaun Baker, Albert T. Corbett, Kenneth R. Koedinger, and Angela Z. Wagner. Off-task behavior in the cognitive tutor classroom: when students "game the system". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 383–390, Vienna, Austria, April 2004. ACM.

[20] Ntombikayise Banda and Peter Robinson. Multimodal affect recognition in intelligent tutoring systems. In *Proceedings of the 5th International Conference of Affective Computing and Intelligent Interaction (ACII)*, pages 200–207, Berlin, Heidelberg, 2011. Springer.

[21] Xuan Bao, Songchun Fan, Alexander Varshavsky, Kevin Li, and Romit Roy Choudhury. Your reactions suggest you liked the movie: Automatic content rating via reaction sensing. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 197–206, Zurich, Switzerland, September 2013. Association for Computing Machinery.

[22] Richard G. Baraniuk. Compressive sensing [lecture notes]. *IEEE Signal Processing Magazine*, 24 (4):118–121, July 2007.

[23] Shaowen Bardzell, Jeffrey Bardzell, and Tyler Pace. Understanding affective interaction: emotion, engagement, and internet videos. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction (ACII)*, 40, pages 1–8, Amsterdam, Netherlands, September 2009. IEEE.

[24] Richard Bartle. Hearts, clubs, siamonds, apades: Players who suit MUDs. *The Journal of Virtual Environments*, 1(1), 1996.

[25] Gian-Marco Baschera, Alberto Giovanni Busetto, Severin Klingler, Joachim M. Buhmann, and Markus Gross. Modeling engagement dynamics in spelling learning. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education (AIED)*, pages 31–38, Auckland, New Zealand, June 2011. Springer.

[26] Marek Bell, Matthew Chalmers, L. Fontaine, Matthew Higgs, Alistair Morrison, John Rooksby, Mattias Rost, and Scott Sherwood. Experiences in logging everyday app use. In *Proceedings of the Digital Economy (DE)*, pages 4–6, Salford, UK, 2013. ACM.

[27] Lisa Benson, Dean Elliott, Michael Grant, Doug Holschuh, Beaumie Kim, Hyeonjin Kim, Erick Lauber, Sebastian Loh, and Thomas C. Reeves. Usability and instructional design heuristics for e-learning evaluation. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA)*, pages 1615–1621, Denver, Colorado, 2002. Association for the Advancement of Computing in Education (AACE).

[28] Niels Ole Berntsen and Laila Dybkjær. Usability data analysis and evaluation. In *Multimodal Usability*, chapter 16, pages 351–385. Springer-Verlag, London, September 2010.

[29] Nagella Uday Bhaskar and P. Govindarajulu. Context exploration for requirements elicitation in mobile learning application development. *International Journal of Computer Science and Network Security (IJCSNS)*, 8(8):292–299, 2008.

[30] Ariel Bleicher. Rise of the Eye Phones [News]. *IEEE Spectrum*, 50(5):9–10, May 2013.

[31] Herbert Bless, Gerald L. Clore, Norbert Schwarz, Verena Golisano, Christina Rabe, and Marcus Wölk. Mood and the use of scripts: Does a happy mood really lead to mindlessness? *Journal of Personality and Social Psychology*, 71(4):665–679, 1996.

[32] Paulo Blikstein and Marcelo Worsley. Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2):220–238, September 2016.

[33] Paris Mavromoustakos Blom, Sander Bakkes, Chek Tien Tan, Shimon Whiteson, Diederik Roijers, Roberto Valenti, and Theo Gevers. Towards personalised gaming via facial expression recognition. In *Proceedings of the Tenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, pages 30–36, Raleigh, NC, October 2014. AAAI Press.

[34] Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger, and Gernot Bauer. Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*, pages 47–56, Stockholm, Sweden, August 2011. Association for Computing Machinery.

[35] Nigel Bosch. Multimodal affect detection in the wild: Accuracy, availability, and generalizability. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI)*, pages 645–649, Seattle, Washington, USA, November 2015. Association for Computing Machinery.

[36] Nigel Bosch, Huili Chen, Sidney D'Mello, Ryan Baker, and Valerie Shute. Accuracy vs. availability heuristic in multimodal affect detection in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI)*, pages 267–274, Seattle, Washington, USA, November 2015. Association for Computing Machinery.

[37] LaVonda Brown and Ayanna M. Howard. Assessment of engagement for intelligent educational agents: A pilot study with middle school students. *ASEE Computers in Education (CoED) Journal*, 5(4):96–106, 2014.

[38] Patrick Brundell, Paul Tennent, Chris Greenhalgh, Dawn Knight, Andy Crabtree, Claire O'Malley, Shaaron Ainsworth, David Clarke, Ronald Carter, and Svenja Adolphs. Digital Replay System (DRS): a tool for interaction analysis. In *Proceedings of the International Conference on Learning Sciences (ICLS) - Workshop on Interaction Analysis*, Utrecht, The Netherlands, 2008.

[39] Susan Bull. Supporting learning with Open Learner Models. In *Proceedings of the 4th Hellenic Conference on Information and Communication Technologies in Education (ICT in Education)*, pages 47–61, Athens, Greece, September 2004. Hellenic Association of ICT in Education (HAICTE).

[40] Susan Bull and Judy Kay. Open Learner Models. In Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi, editors, *Advances in Intelligent Tutoring Systems*, pages 301–322. Springer, Berlin, Heidelberg, 2010.

[41] Paolo Burelli, Georgios Triantafyllidis, and Ioannis Patras. Non-Invasive player experience estimation from body motion and game context. In *Proceedings of the IEEE Conference on Computational Intelligence and Games (CIG)*, 68, pages 1–7, Dortmund, Germany, 2014. IEEE.

[42] JaeHwan Byun, Christian S. Loh, and Ting Zhou. Assessing play-learners' performance in serious game environments by using in situ data: Using eye tracking for serious game analytics. In *Paper presented at the Annual Conference of the Association for Educational Communications and Technology (AECT)*, Jacksonville, FL, 2014. Research and Theory Division Jacksonville, FL, aect.org.

[43] Patrick Carmichael and Katy Jordan. Semantic web technologies for education - time for a "turn to practice"? *Technology, Pedagogy and Education*, 21(2):153–169, 2012.

[44] Keng-hao Chang, Drew Fisher, John Canny, and Björn Hartmann. How's my mood and stress? an efficient speech analysis library for unobtrusive monitoring on mobile phones. In *Proceedings of the 6th International Conference on Body Area Networks*, pages 71–77, Beijing, China, November 2011. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[45] Maher Chaouachi, Imane Jraidi, and Claude Frasson. MENTOR: A physiologically controlled tutoring system. In *Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 56–67, Cham, 2015. Springer International Publishing.

[46] Kyoung-Sic Cho, In-Ho Choi, and Yong-Guk Kim. Robust facial expression recognition using a smartphone working against illumination variation. *Applied Mathematics & Information Sciences*, 6(2S):403S–408S, 2012.

[47] M. Chuah and M. Diblasio. Smartphone based Autism social alert system. In *Proceedings of the 8th International Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, pages 6–13, Chengdu, China, December 2012. IEEE.

[48] Meng-Che Chuang, Raja Bala, Edgar A. Bernal, Peter Paul, and Aaron Burry. Estimating gaze direction of vehicle drivers using a smartphone camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 165–170, Columbus, OH, June 2014. IEEE.

[49] Gregory K. W. K. Chung. Guidelines for the design and implementation of game telemetry for serious games analytics. In Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler, editors, *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, pages 59–79. Springer International Publishing, Cham, 2015.

[50] Gregory K. W. K. Chung and Deirdre S. Kerr. A primer on data logging to support extraction of meaningful information from educational games: An example from save patch. Technical Report CRESST Report 814, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), California, LA, March 2012.

[51] Cristina Conati. How to evaluate models of user affect? In *Affective Dialogue Systems, Tutorial and Research Workshop (ADS)*, chapter 30, pages 288–300. Springer, Kloster Irsee, Germany, 2004.

[52] Cristina Conati and Mirela Gutica. Interaction with an edu-game: A detailed analysis of student emotions and judges' perceptions. *International Journal of Artificial Intelligence in Education (IJAIED)*, 26(4):975–1010, December 2016.

[53] Cristina Conati and Heather Maclaren. Modeling user affect from causes and effects. In *Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 4–15, Berlin, Germany, 2009. Springer.

[54] Adam Cooper. Learning analytics and interoperability - The big picture in brief. Technical report, Learning Analytics Community Exchange (LACE), March 2014. Project Description.

[55] Ben Cowley, Niklas Ravaja, and Tuija Heikura. Cardiovascular physiology predicts learning effects in a serious game activity. *Computers & Education*, 60(1):299–309, January 2013.

[56] Benjamin Cowley and Niklas Ravaja. Learning in balance: Using oscillatory EEG biomarkers of attention, motivation and vigilance to interpret game-based learning. *Cogent Education*, 1(1): 962236, December 2014.

[57] Mark Cox, Jesus Nuevo-Chiquero, Jason Saragih, and Simon Lucey. CSIRO Face Analysis SDK. Technical report, AFGR, Brisbane, Australia, 2013.

[58] Heather Desurvire and Charlotte Wiberg. Game usability heuristics (PLAY) for evaluating and designing better games: The next iteration. In *Proceedings of the Third International Conference on Online Communities and Social Computing (OCSC)*, pages 557–566, Berlin, Heidelberg, 2009. Springer.

[59] Heather Desurvire, Martin Caplan, and Jozsef A. Toth. Using heuristics to evaluate the playability of games. In *Proceedings of the ACM CHI Annual Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, pages 1509–1512, Vienna, Austria, April 2004. Association for Computing Machinery.

[60] Beth Dietz-Uhler and Janet E. Hurn. Using learning analytics to predict (and improve) student success: A faculty perspective. *Journal of Interactive Online Learning*, 12(1):17–26, 2013.

[61] Priyesh N. Dixit and G. Michael Youngblood. Understanding information observation in interactive 3D environments. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games (Sandbox)*, pages 163–170, Los Angeles, California, August 2008. Association for Computing Machinery.

[62] Anders Drachen, Lennart E. Nacke, Georgios Yannakakis, and Anja Lee Pedersen. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games (Sandbox)*, pages 49–54, Los Angeles, California, July 2010. Association for Computing Machinery.

[63] Brandon Drenikow and Pejman Mirza-Babaei. Vixen: Interactive visualization of gameplay experiences. In *Proceedings of the 12th International Conference on the Foundations of Digital Games (FDG)*, 3, pages 1–10, Hyannis, MA, August 2017. Association for Computing Machinery.

[64] Erik Duval. Attention Please! Learning analytics for visualization and recommendation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge (LAK)*, pages 9–17, Banff, Alberta, Canada, February 2011. Association for Computing Machinery.

[65] Anna Lea Dyckhoff, Dennis Zielke, Mareike Bültmann, Mohamed Amine Chatti, and Ulrik Schroeder. Design and implementation of a learning analytics toolkit for teachers. *Educational Technology & Society*, 15(3):58–76, 2012.

[66] Gregory Dyke, Kristine Lund, and Jean-Jacques Girardot. Tatiana: An environment to support the CSCL analysis process. In *Proceedings of the 9th international conference on Computer supported collaborative learning - Volume 1 (CSCL)*, pages 58–67, Rhodes, Greece, June 2009. International Society of the Learning Sciences.

[67] Anastasios A. Economides. Adaptive context-aware pervasive and ubiquitous learning. *International Journal of Technology Enhanced Learning (IJTEL)*, 1(3):169–192, May 2009.

[68] Paul Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, 19:207–283, 1971.

[69] Rana el Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In Branislav Kisačanin, Vladimir Pavlović, and Thomas S. Huang, editors, *Real-Time Vision for Human-Computer Interaction*, pages 181–200. Springer US, Boston, MA, 2005.

[70] Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa. *Game analytics: Maximizing the value of player data*. Springer-Verlag, London, 2013.

[71] Clayton Epp, Michael Lippold, and Regan L. Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 715–724, Vancouver, BC, Canada, May 2011. Association for Computing Machinery.

[72] Aysu Ezen-Can, Joseph F. Grafsgaard, James C. Lester, and Kristy Elizabeth Boyer. Classifying student dialogue acts with multimodal learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK)*, pages 280–289, Poughkeepsie, New York, March 2015. Association for Computing Machinery.

[73] Jean-Claude Falmagne, Dietrich Albert, Christopher Doble, David Eppstein, and Xiangen Hu. *Knowledge spaces: Applications in education*. Springer, Berlin, Heidelberg, 2013.

[74] Ian Timothy Fischer, Cynthia Kuo, Ling Huang, and Mario Frank. Smartphones: Not smart enough? In *Proceedings of the second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices (SPSM)*, pages 27–32, Raleigh, North Carolina, USA, October 2012. Association for Computing Machinery.

[75] Adam Fouse, Nadir Weibel, Edwin Hutchins, and James D. Hollan. ChronoViz: a system for supporting navigation of time-coded data. In *Proceedings of the ACM CHI Annual Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, pages 299–304, Vancouver, BC, Canada, May 2011. Association for Computing Machinery.

[76] Sara Freitas, Genaro Rebolledo-Mendez, Fotis Liarokapis, George Magoulas, and Alexandra Poulovassilis. Learning as immersive experiences: Using the Four-dimensional framework for designing and evaluating immersive learning experiences in a virtual world. *British Journal of Educational Technology*, 41(1):69–85, January 2010.

[77] Jon Froehlich, Mike Y. Chen, Sunny Consolvo, Beverly Harrison, and James A. Landay. MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services (MobiSys)*, pages 57–70, San Juan, Puerto Rico, June 2007. Association for Computing Machinery.

[78] Fong-Ling Fu and Sheng-Chin Yu. Three layered thinking model for designing web-based educational games. In *Proceedings of the 17th International Conference on Web-Based Learning: Advances in Web Based Learning (ICWL)*, pages 265–274, Berlin, Heidelberg, 2008. Springer.

[79] Rosemary Garris, Robert Ahlers, and James E. Driskell. Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, 33(4):441–467, December 2002.

[80] Sari Gilbert. *Designing gamified systems: Meaningful play in interactive entertainment, marketing and education*. Routledge, New York, September 2015.

[81] Stefan Göbel, André De, Carvalho Rodrigues, Florian Mehm, and Ralf Steinmetz. Narrative game-based learning objects for story-based digital educational games. In *Proceedings of the 1st International Open Workshop on Intelligent Personalization and Adaptation in Digital Educational Games*, 14, pages 43–53, Graz, January 2009.

[82] Stefan Göbel, Viktor Wendel, Christopher Ritter, and Ralf Steinmetz. Personalized, adaptive digital educational games using narrative game-based learning objects. In *Entertainment for Education. Digital Techniques and Systems*, pages 438–445, Berlin, Heidelberg, 2010. Springer.

[83] Stefan Göbel, Michael Gutjahr, and Sandro Hardy. Evaluation of serious games. In Klaus Bredl and Wolfgang Bösche, editors, *Serious Games and Virtual Worlds in Education, Professional Development, and Healthcare*, chapter 7, pages 105–115. IGI Global, 2013.

[84] July Silveira Gomes, Mohamed Yassine, Marcelo Worsley, and Paulo Blikstein. Analysing engineering expertise of high school students using eye tracking and multimodal learning analytics. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM)*, pages 375–377. International Educational Data Mining Society, January 2013.

[85] Sergio Gómez, Panagiotis Zervas, Demetrios G. Sampson, and Ramón Fabregat. Context-aware adaptive and personalized mobile learning delivery supported by UoLmP. *Journal of King Saud University Computer and Information Sciences*, 26(1):47–61, January 2014.

[86] Arthur Graesser, Patrick Chipman, Frank Leeming, and Suzanne Biedenbach. Deep learning and emotion in serious games. In Peter Vorderer Ute Ritterfeld, Michael Cody, editor, *Serious Games: Mechanisms and Effects*, chapter 6, pages 81–100. Routledge, Taylor & Francis New York and London, 2009.

[87] Joseph F. Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester II. Embodied affect in tutorial dialogue. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED)*, 1, pages 1–10. Springer, 2013.

[88] Joseph F. Grafsgaard, Joseph B. Wiggins, Alexandria Katarina Vail, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI)*, pages 42–49, Istanbul, Turkey, November 2014. Association for Computing Machinery.

[89] Joseph F. Grafsgaard, Seung Y. Lee, Bradford W. Mott, Kristy Elizabeth Boyer, and James C. Lester. Modeling self-efficacy across age groups with automatically tracked facial expression. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, pages 582–585, Cham, 2015. Springer International Publishing.

[90] Dan J. Graham and J. Aaron Hipp. Emerging technologies to promote and evaluate physical activity: Cutting-edge research and future directions. *Frontiers in Public Health*, 2:66, 2014.

[91] Mirela Gutica and Cristina Conati. Student emotions with an edu-game: A detailed analysis. In *Proceedings of the Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 534–539, Geneva, September 2013. IEEE.

[92] Eun Y. Ha, Jonathan P. Rowe, Bradford W. Mott, and James C. Lester. Recognizing player goals in open-ended digital games with Markov Logic networks. In Gita Sukthankar, Christopher Geib, Hung Hai Bui, David Pynadath, and Robert P. Goldman, editors, *Plan, Activity, and Intent Recognition: Theory and Practice*, pages 289–311. Morgan Kaufmann, Amsterdam ; Boston, February 2014.

[93] Richard Halverson and V. Elizabeth Owen. Game-based assessment: an integrated model for capturing evidence of learning in play. *International Journal of Learning Technology (IJLT)*, 9(2): 111–138, August 2014.

[94] Seungyeop Han, Rajalakshmi Nandakumar, Matthai Philipose, Arvind Krishnamurthy, and David Wetherall. GlimpseData: Towards continuous vision-based personal analytics. In *Proceedings of the 2014 Workshop on Physical Analytics (ICMI)*, pages 31–36, Bretton Woods, New Hampshire, USA, June 2014. Association for Computing Machinery.

[95] Erik Harpstead, Brad A. Myers, and Vincent Aleven. In search of learning: Facilitating data analysis in educational games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 79–88, Paris, France, April 2013. Association for Computing Machinery.

[96] Erik Harpstead, Christopher J. MacLellan, Vincent Aleven, and Brad A. Myers. Replay analysis in open-ended educational games. In Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler, editors, *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, pages 381–399. Springer International Publishing, Cham, 2015.

[97] Jannicke Baalsrud Hauge, Riccardo Berta, Giusy Fiucci, Baltasar Fernández Manjón, Carmen Padrón-Nápoles, Wim Westra, and Rob Nadolski. Implications of learning analytics for serious game design. In *Proceedings of the IEEE 14th International Conference on Advanced Learning Technologies ((ICALT))*, pages 230–232. IEEE, July 2014.

[98] Jannicke Balsrud Hauge, Ioana Andreea Stanescu, Antoniu Stefan, Theodore Lim, Maira B. Carvalho, Sandy Louchart, and Sylvester Arnab. Serious game mechanics and opportunities for reuse. In *Proceedings of the 11th International Conference eLearning and Software for Education (eLSE)*, volume 2, pages 19–27, Bucharest, 2015. Carol I NDU Publishing House.

[99] Yuko Hirabe, Yutaka Arakawa, and Keiichi Yasumoto. Logging all the touch operations on Android. In *Proceedings of the Seventh International Conference on Mobile Computing and Ubiquitous Networking (ICMU)*, pages 93–94, Singapore, January 2014. IEEE.

[100] Camillan Huang. Designing high-quality interactive multimedia learning modules. *Computerized Medical Imaging and Graphics*, 29(2):223–233, March 2005.

[101] Eva Hudlicka. Affective game engines: Motivation and requirements. In *Proceedings of the 4th international conference on foundations of digital games*, pages 299–306, Orlando, Florida, USA, January 2009. ACM.

[102] William Huitt. Bloom et al.'s taxonomy of the cognitive domain. *Educational Psychology Interactive*, 22, 2004.

[103] Ioanna Iacovides and Anna L. Cox. Moving beyond fun: Evaluating serious experience in digital games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, pages 2245–2254, New York, USA, April 2015. ACM.

[104] Curtis S. Ikehara, Jiecai He, and Martha E. Crosby. Issues in Implementing augmented cognition and gamification on a mobile platform. In Dylan D. Schmorrow and Cali M. Fidopiastis, editors, *Foundations of Augmented Cognition*, pages 685–694. Springer, Berlin, Heidelberg, 2013.

[105] Klaus P. Jantke. Patterns of game playing behavior as indicators of mastery. In Dirk Ifenthaler, Deniz Eseryel, and Xun Ge, editors, *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, pages 85–103. Springer, New York, NY, 2012.

[106] Larry Johnson, Becker S. Adams, Malcolm Cummins, Victoria Estrada, Alex Freeman, and Holly Ludgate. *The NMC Horizon Report: 2013 Higher Education Edition*. New Media Consortium, Austin, TX, 2013.

[107] Michael Johnston and Srinivas Bangalore. Finite-state multimodal integration and understanding. *Natural Language Engineering*, 11(2):159–187, June 2005.

[108] Simon Joslin, Ross Brown, and Penny Drennan. The gameplay visualization manifesto: a framework for logging and visualization of online gameplay data. *Computers in Entertainment*, 5(3):6, July 2007.

[109] Anne Joutsenvirta, A. Kekäläinen, M. Cankar, Titti Kallio, and Anu Kankainen. Usability testing of mobile applications: A comparison between laboratory and field testing. *Journal of Usability Studies*, 1(1):4–16, January 2005.

[110] Shiming Kai, Luc Paquette, Ryan S. Baker, Philip N. Bosch, Sidney K. D'Mello, Jaclyn Ocumpaugh, Valerie J. Shute, and Matthew Ventura. A comparison of face-based and interaction-based affect detectors in physics playground. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM)*, pages 77–84, Madrid, Spain, 2015. EDM.

[111] Vani Kalloo, Kinshuk, and Permanand Mohan. Personalized game based mobile learning to assist high school students with mathematics. In *Proceedings of the 10th IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 485–487, Sousse, Tunisia, July 2010. IEEE.

[112] Vaiva Kalnikaite, Abigail Sellen, Steve Whittaker, and David Kirk. Now let me see where I was: Understanding how lifelogs mediate memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, volume 3, pages 2045–2054, Atlanta, Georgia, June 2010. ACM.

[113] Ruth Kanfer and Phillip Ackerman. Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology*, 74(4):657–690, August 1989.

[114] Seungwoo Kang, Sungjun Kwon, Chungkuk Yoo, Sangwon Seo, Kwang Park, Junehwa Song, and Youngki Lee. Sinabro: Opportunistic and unobtrusive mobile ECG monitoring system. In *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications (HotMobile )*, 11, pages 1–6, New York, NY, February 2014. ACM.

[115] Deirdre Kerr and Gregory K. W. K. Chung. Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining (JEDM)*, 4(1):144–182, October 2012.

[116] Deirdre S. Kerr and Gregory K. W. K. Chung. Using cluster analysis to extend usability testing to instructional content. Technical Report CRESST Report 816, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), May 2012.

[117] Myint Swe Khine. *Learning to play: Exploring the future of education with video games*. Peter Lang Inc., New York, November 2010.

[118] Michael D. Kickmeier-Rust. Talking digital educational games. In *Proceedings of the 1st International Open Workshop on Intelligent Personalization and Adaptation in Digital Educational Games*, pages 55–66. Springer, 2009.

[119] Michael D. Kickmeier-Rust and Dietrich Albert. *An alien's guide to multi-adaptive educational computer games*. Informing Science Press, February 2012.

[120] Michael D. Kickmeier-Rust and Dietrich Albert. Learning Analytics to support the use of virtual worlds in the classroom. In *Proceedings of the International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data (HCI-KDD)*, pages 358–365, Berlin, Heidelberg, 2013. Springer.

[121] Kristian Kiili, Harri Ketamo, and Michael D. Kickmeier-Rust. Eye tracking in game-based learning research and game design. *International Journal of Serious Games (IJSG)*, 1(2):51–65, 2014.

[122] Kristian Kiili, Harri Ketamo, and Michael D. Kickmeier-Rust. Evaluating the usefulness of eye tracking in game-based learning. *International Journal of Serious Games (IJSG)*, 8(2), 2014.

[123] Kristian Kiili, Timo Lainema, Sara de Freitas, and Sylvester Arnab. Flow framework for analyzing the quality of educational games. *Entertainment Computing*, 5(4):367–377, December 2014.

[124] Jun H. Kim, Daniel V. Gunn, Eric Schuh, Bruce Phillips, Randy J. Pagulayan, and Dennis Wixon. Tracking real-time user experience (TRUE): a Comprehensive instrumentation solution for complex systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 443–452, Florence, Italy, April 2008. Association for Computing Machinery.

[125] Kenneth R. Koedinger, Ryan S. J. d. Baker, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John Stamper. A data repository for the EDM community: The PSLC dataShop. In Cristobal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan S. J. d Baker, editors, *Handbook of Educational Data Mining*, pages 43–55. CRC Press, Boca Raton, October 2010.

[126] Johannes Konert, Kristina Richter, Florian Mehm, Stefan Göbel, Regina Bruder, and Ralf Steinmetz. PEDALE – A peer education diagnostic and learning environment. *Journal of Educational Technology & Society*, 15(4):27–38, 2012.

[127] Marguerite L. Koole. Practical issues in mobile education. In *Proceedings of the Fourth IEEE International Workshop on Wireless, Mobile and Ubiquitous Technology in Education (WMTE)*, pages 142–146, USA, November 2006. IEEE Computer Society.

[128] Klaus Korossy. Modeling knowledge as competence and performance. In Dietrich Albert and Josef Lukas, editors, *Knowledge Spaces: Theories, Empirical Research, and Applications*, pages 103–132. Psychology Press, New York, NY, August 1999.

[129] Kai Kunze, Masakazu Iwamura, Koichi Kise, Seiichi Uchida, and Shinichiro Omachi. Activity recognition for the mind: Toward a cognitive "Quantified Self". *Computer (IEEE)*, 46(10):105–108, October 2013.

[130] Tracey L. Leacock and John C. Nesbit. A framework for evaluating the quality of multimedia learning resources. *Journal of Educational Technology & Society*, 10(2):44–59, 2007.

[131] Hosub Lee, Young Sang Choi, Sunjae Lee, and I. P. Park. Towards unobtrusive emotion recognition for affective social communication. In *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC)*, pages 260–264, Las Vegas, NV, January 2012. IEEE.

[132] James Lester, Bradford W. Mott, Jonathan P. Rowe, and Jennifer Sabourin. Learner modeling to predict real-time affect in serious games. *Design Recommendations for Intelligent Tutoring Systems*, 1:199–208, 2013.

[133] Susan M. Letourneau. Intensive data at CREATE - Research approaches & future directions. In *Proceedings of the Simulations and Games Learning Analytics and Data Mining Workshop*, 2013.

[134] Timo Leuders, Regina Bruder, Ulf Kroehne, Dominik Naccarella, Renate Nitsch, Jan Henning-Kahmann, Augustin Kelava, and Markus Wirtz. Development, validation, and application of a competence model for Mathematical problem solving by using and translating representations of functions. In Detlev Leutner, Jens Fleischer, Juliane Grünkorn, and Eckhard Klieme, editors, *Competence Assessment in Education: Research, Models and Instruments*, pages 389–406. Springer International Publishing, Cham, 2017.

[135] Roy Levy. Dynamic Bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, 54(6):771–794, November 2019.

[136] Jian-Sheng Li, Yun-Xia Han, and Shan Liu. Physiological evaluation of the players' emotions in different educational games. *American Journal of Educational Research*, 2(9):735–739, January 2014.

[137] Zhung-Xun Liao, Yi-Chin Pan, Wen-Chih Peng, and Po-Ruey Lei. On mining mobile apps usage behavior for predicting apps usage in smartphones. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM)*, pages 609–618, San Francisco, California, USA, October 2013. Association for Computing Machinery.

[138] Florian Lingenfelser, Johannes Wagner, Elisabeth André, Gary McKeown, and Will Curran. An event driven fusion approach for enjoyment recognition in real-time. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM)*, pages 377–386, Orlando, Florida, USA, November 2014. Association for Computing Machinery.

[139] Christian S. Loh. Information Trails: In-process assessment of game-based learning. In Dirk Ifenthaler, Deniz Eseryel, and Xun Ge, editors, *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, pages 123–144. Springer, New York, NY, 2012.

[140] Christian Sebastian Loh and Yanyan Sheng. Maximum Similarity Index (MSI): A metric to differentiate the performance of novices vs. multiple-experts in serious games. *Computers in Human Behavior*, 39:322–330, October 2014.

[141] Christian Sebastian Loh and Yanyan Sheng. Measuring expert performance for serious games analytics: From data to insights. In Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler, editors, *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, pages 101–134. Springer International Publishing, Cham, 2015.

[142] Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler. Serious games analytics: Theoretical framework. In Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler, editors, *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, pages 3–29. Springer International Publishing, Cham, 2015.

[143] Hong Lu, Jun Yang, Zhigang Liu, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. The Jigsaw continuous sensing engine for mobile phone applications. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 71–84, Zürich, Switzerland, November 2010. Association for Computing Machinery.

[144] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp)*, pages 351–360, Pittsburgh, Pennsylvania, September 2012. Association for Computing Machinery.

[145] Weiquan Lu, Chenchen Sun, Timo Bleeker, Yingdan You, Shintaro Kitazawa, and Ellen Yi-Luen Do. Sensorendipity: a real-time web-enabled smartphone sensor platform for idea generation and prototyping. In *Proceedings of the Second International Symposium of Chinese CHI*, pages 11–18, Toronto, Ontario, Canada, April 2014. Association for Computing Machinery.

[146] Simon Lucey, Ahmed Bilal Ashraf, and Jeffrey F. Cohn. Investigating spontaneous facial action recognition through AAM representations of the face. In Kresimir Delac and Mislav Grgic, editors, *Face Recognition*, chapter 14, pages 275–286. IntechOpen, July 2007.

[147] Leah P. Macfadyen and Shane Dawson. Mining LMS data to develop an early warning system for educators: A proof of concept. *Computers & Education*, 54(2):588–599, February 2010.

[148] Thomas W. Malone. What makes things fun to learn? Heuristics for designing instructional computer games. In *Proceedings of the 3rd ACM SIGSMALL Symposium and the first SIGPC Symposium on Small Systems*, pages 162–169, Palo Alto, California, USA, September 1980. Association for Computing Machinery.

[149] Marco Manca, Fabio Paternò, and Carmen Santoro. Adaptive mobile multimodal user interfaces for dynamic accessibility. In *Proceedings of the ACM CHI Annual Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, ACM, April 2013. ACM.

[150] Regan L. Mandryk and M. Stella Atkins. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, 65(4):329–347, April 2007.

[151] Kunal Mankodiya, Vinod Sharma, Rolando Martins, Ishan Pande, Samay Jain, Neal Ryan, and Rajeev Gandhi. Understanding user's emotional engagement to the contents on a smartphone display: Psychiatric prospective. In *Proceedings of the 2013 IEEE 10th International Conference on Ubiquitous Intelligence & Computing and 2013 IEEE 10th International Conference on Autonomic & Trusted Computing (UIC-ATC)*, pages 631–637, USA, December 2013. IEEE Computer Society.

[152] Raphael Marczak and Gareth R. Schott. Audio visual analysis of player experience: Feedback-based gameplay metrics. In Petri Lankoski and Staffan Bjœrk, editors, *Game Research Methods*, pages 207–230. ETC Press, Pittsburgh, PA, USA, January 2015.

[153] Raphael Marczak, Jasper van Vught, Gareth Schott, and Lennart E. Nacke. Feedback-based gameplay metrics: Measuring player experience via automatic visual analysis. In *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System (IE)*, 6, pages 1–10, Auckland, New Zealand, July 2012. Association for Computing Machinery.

[154] Panos Markopoulos, Janet C. Read, Stuart MacFarlane, and Johanna Hoysniemi. *Evaluating children's interactive products: Principles and practices for interaction designers*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[155] Bertrand Marne, John Wisdom, Benjamin Huynh-Kim-Bang, and Jean-Marc Labat. The six facets of serious game design: A methodology enhanced by our design pattern library. In *Proceedings of the European Conference on Technology Enhanced Learning (ECTEL)*, pages 208–221, Berlin, Heidelberg, 2012. Springer.

[156] Nurul Hidayah Mat Zain, Fariza Hanis Abdul Razak, Azizah Jaafar, and Mohd Firdaus Zulkipli. Eye tracking in educational games environment: Evaluating user interface design through Eye tracking patterns. In Halimah Badioze Zaman, Peter Robinson, Maria Petrou, Patrick Olivier, Timothy K. Shih, Sergio Velastin, and Ingela Nyström, editors, *Visual Informatics: Sustaining Research and Innovations*, pages 64–73. Springer, Berlin, Heidelberg, 2011.

[157] J. Matias Kivikangas, Lennart Nacke, and Niklas Ravaja. Developing a triangulation system for digital game events, observational video, and psychophysiological data to study emotional responses to a virtual character. *Entertainment Computing*, 2(1):11–16, January 2011.

[158] Igor Mayer. Towards a comprehensive methodology for the research and evaluation of serious games. *Procedia Computer Science*, 15:233–247, 2012.

[159] Igor Mayer, Geertje Bekebrede, Casper Harteveld, Harald Warmelink, Qiqi Zhou, Theo van Ruijven, Julia Lo, Rens Kortmann, and Ivo Wenzler. The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology*, 45(3):502–507, June 2014.

[160] Daniel McDuff, Rana El Kaliouby, Thibaud Senechal, David Demirdjian, and Rosalind Picard. Automatic measurement of ad preferences from facial responses gathered over the internet. *Image and Vision Computing*, 32(10):630–640, October 2014.

[161] Lorna McKnight and Janet C. Read. PLU-E: a proposed framework for planning and conducting evaluation studies with children. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction (BCS-HCI)*, pages 126–131, Newcastle-upon-Tyne, United Kingdom, July 2011. BCS Learning & Development Ltd.

[162] Mark Mcmahon and Cesar Ojeda. A model of immersion to guide the design of serious games. In *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education E-Learn*, pages 1833–1842, Chesapeake, VA, January 2008. AACE.

[163] Donald McMillan, Alistair Morrison, and Matthew Chalmers. Categorised ethical guidelines for large scale mobile HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1853–1862, Paris, France, April 2013. Association for Computing Machinery.

[164] Ben Medler, Michæl John, and Jeff Lane. Data Cracker: Developing a visual game analytic tool for analyzing online gameplay. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2365–2374, Vancouver, BC, Canada, May 2011. Association for Computing Machinery.

[165] Florian Mehm, Viktor Wendel, Stefan Göbel, and Ralf Steinmetz. Bat Cave: A testing and evaluation platform for digital educational games. In Bente Meyer, editor, *ECGBL2009- 4th European Conference on Games-Based Learning: ECGBL 2009*, pages 251–260. Academic Conferences Limited, Reading, England, December 2010.

[166] Florian Mehm, Stefan Göbel, and Ralf Steinmetz. Authoring of serious adventure games in StoryTec. In Stefan Göbel, Wolfgang Müller, Bodo Urban, and Josef Wiemeyer, editors, *E-Learning and Games for Training, Education, Health and Sports*, pages 144–154. Springer, Berlin, Heidelberg, 2012.

[167] Ricardo Javier Rademacher Mena. Game assesement using the E/E Grid. In Leonard Annetta and Stephen C. Bronack, editors, *Serious Educational Game Assessment: Practical Methods and Models for Educational Games, Simulations and Virtual Worlds*, pages 95–117. SensePublishers, Rotterdam, 2011.

[168] Debra K. Meyer and Julianne C. Turner. Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review*, 18(4):377–390, December 2006.

[169] Emiliano Miluzzo, Nicholas D. Lane, Shane B. Eisenman, and Andrew T. Campbell. CenceMe : Injecting sensing presence into social networking applications. In *Proceedings of the European Conference on Smart Sensing and Context ( EuroSSC )*, 1, pages 1–28, Berlin, Heidelberg, 2007. Springer.

[170] Miroslav Minović and Miloš Milovanović. Real-time learning analytics in educational games. In *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality (TEEM)*, pages 245–251, Salamanca, Spain, November 2013. Association for Computing Machinery.

[171] Pejman Mirza-Babaei, Günter Wallner, Graham McAllister, and Lennart E. Nacke. Unified visualization of quantitative and qualitative playtesting data. In *Proceedings of the ACM CHI Annual Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, pages 1363–1368. ACM, 2014.

[172] Robert J. Mislevy, Russell G. Almond, and Janice F. Lukas. A brief introduction to evidence-centered design. Technical Report Research Report 03-16, Educational Testing Service, Princeton, NJ, 2003.

[173] Konstantin Mitgutsch and Narda Alvarado. Purposeful by Design? a serious game design assessment framework. In *Proceedings of the International Conference on the Foundations of Digital Games (FDG)*, pages 121–128, Raleigh, North Carolina, May 2012. Association for Computing Machinery.

[174] Antonija Mitrovic and Brent Martin. Evaluating the effect of open student models on self-assessment. *International Journal of Artificial Intelligence in Education*, 17(2):121–144, 2007.

[175] Pablo Moreno-Ger, Javier Torrente, Yichuan Grace Hsieh, and William T. Lester. Usability testing for serious games: Making informed design decisions with user data. *Advances in Human-Computer Interaction (Hindawi)*, 2012:13–26, January 2012.

[176] Allistair Morrison, Paul Tennent, and Matthew Chalmers. Coordinated visualisation of video and system log data. In *Proceedings of the Fourth International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV)*, pages 91–102, USA, July 2006. IEEE Computer Society.

[177] Allistair Morrison, Paul Tennent, and Matthew Chalmers. Coordinated visualisation of video and system log data. In *Proceedings of the Fourth International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV)*, pages 91–102, USA, July 2006. IEEE Computer Society.

[178] Theano Moussouri and George Roussos. Mobile Sensing, BYOD and Big Data Analytics: New technologies for audience research in museums. *Participations: Journal of Audience and Reception Studies*, 11(1):270–285, 2014.

[179] Stephan Mueller, Barbara Solenthaler, Mubbasir Kapadia, Seth Frey, Severin Klingler, Richard P. Mann, Robert W. Sumner, and Markus Gross. HeapCraft: Interactive data exploration and visualization tools for understanding and influencing player behavior in minecraft. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games (MIG)*, pages 237–241, Paris, France, November 2015. Association for Computing Machinery.

[180] Mary Muir and Cristina Conati. Understanding student attention to adaptive hints with eye-tracking. In *Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 148–160, Berlin, Heidelberg, 2012. Springer.

[181] Catherine Mulwa, Seamus Lawless, Mary Sharp, Inmaculada Arnedillo-Sanchez, and Vincent Wade. Adaptive educational hypermedia systems in technology enhanced learning: a literature review. In *Proceedings of the 2010 ACM Conference on Information Technology Education (SIGITE)*, pages 73–84, Midland, Michigan, USA, October 2010. Association for Computing Machinery.

[182] Lennart E. Nacke, Anders Drachen, and Stefan Göbel. Methods for evaluating gameplay experience in a serious gaming context. *International Journal of Computer Science in Sport*, 9(2):1–12, 2010.

[183] Lennart E. Nacke, Sophie Stellmach, and Craig A. Lindley. Electroencephalographic assessment of player experience: A pilot study in affective ludology. *Simulation & Gaming*, 42(5):632–655, October 2011.

[184] Lennart E. Nacke, Sophie Stellmach, Dennis Sasse, Joerg Niesenhaus, and Raimund Dachselt. LAIF: A Logging and interaction framework for gaze-based interfaces in virtual entertainment environments. *Entertainment Computing*, 2(4):265–273, January 2011.

[185] Keith Nesbitt, Karen Blackmore, Geoffrey Hookham, Frances Kay-Lambkin, and Peter Walla. Using the Startle Eye-Blink to measure affect in players. In Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler, editors, *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, pages 401–434. Springer International Publishing, Cham, 2015.

[186] Pedro A. Nogueira, Rúben Aguiar, Rui Rodrigues, Eugénio Oliveira, and Lennart E. Nacke. Fuzzy affective player models: a physiology-based hierarchical clustering method. In *Proceedings of the Tenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, pages 132–138, Raleigh, NC, USA, October 2014. AAAI Press.

[187] Pedro A. Nogueira, Vasco Torres, Rui Rodrigues, and Eugénio Oliveira. An annotation tool for automatically triangulating individuals' psychophysiological emotional reactions to digital media stimuli. *Entertainment Computing*, 9-10:19–27, June 2015.

[188] Elena Patricia Nunez Castellar, Kimmo Oksanen, and Jan Van Looy. Assessing game experience: Heart rate variability, in-game behavior and self-report measures. In *Proceedings of The 6th International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 292–296. IEEE, 2015.

[189] Kari Nurmela, Erno Lehtinen, and Tuire Palonen. Evaluating CSCL log files by social network analysis. In *Proceedings of the 1999 Conference on Computer support for Collaborative Learning (CSCL)*, pages 54–es, Palo Alto, California, December 1999. International Society of the Learning Sciences.

[190] Hiroaki Ogata, Mengmeng Li, Bin Hou, Noriko Uosaki, Moushir El-Bishouty, and Yoneo Yano. SCROLL: Supporting to share and reuse ubiquitous learning log in the context of language learning. *Research and Practice in Technology Enhanced Learning*, 6(2):69–82, January 2011.

[191] Paul A. O'Keefe, Susan M. Letourneau, Bruce D. Homer, Ruth N. Schwartz, and Jan L. Plass. Learning from multiple representations: An examination of fixation patterns in a science simulation. *Computers in Human Behavior*, 35:234–242, June 2014.

[192] Tanner Olsen, Katelyn Procci, and Clint Bowers. Serious games usability testing: How to ensure proper usability, playability, and effectiveness. In Aaron Marcus, editor, *Design, User Experience, and Usability. Theory, Methods, Tools and Practice*, pages 625–634. Springer, Berlin, Heidelberg, 2011.

[193] Hasiah Mohamed Omar and Azizah Jaafar. Challenges in the evaluation of educational computer games. In *Proceedings of the International Symposium on Information Technology (ITSim)*, volume 1, pages 11–17, Kuala Lumpur, Malaysia, June 2010. IEEE.

[194] Kerry O'Regan. Emotion and E-learning. *Journal of Asynchronous Learning Network*, 7(3):78–92, September 2003.

[195] Joseph C. Osborn and Michæl Mateas. A game-independent play trace dissimilarity metric. In *Proceedings of the 9th International Conference on the Foundations of Digital Games (FDG)*. Society for the Advancement of the Science of Digital Games, 2014.

[196] Sharon Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, November 1999.

[197] Sharon Oviatt, Adrienne Cohen, and Nadir Weibel. Multimodal learning analytics: Description of math data corpus for ICMI grand challenge workshop. In *Proceedings of the 15th ACM on International conference on multimodal interaction (ICMI)*, pages 563–568, Sydney, Australia, 2013. ACM Press.

[198] Luc Paquette, Jonathan Rowe, Ryan Baker, Bradford Mott, James Lester, Jeanine DeFalco, Keith Brawner, Robert Sottilare, and Vasiliki Georgoulas. Sensor-free or sensor-full: A comparison of data modalities in multi-channel affect Detection. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM)*, pages 93–100, Madrid, Spain, 2016. ERIC.

[199] Pablo Paredes, David Sun, and John Canny. Sensor-less sensing for affective computing and stress management technology. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 459–463, Venice, Italy, May 2013. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[200] Georgios Patsis, Hichem Sahli, Werner Verhelst, and Olga De Troyer. Evaluation of attention levels in a Tetris game using a brain computer interface. In Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli, and Giovanni Semeraro, editors, *User Modeling, Adaptation, and Personalization*, pages 127–138. Springer, Berlin, Heidelberg, 2013.

[201] Reinhard Pekrun. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4): 315–341, December 2006.

[202] Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P. Perry. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, 37(2):91–105, June 2002.

[203] Reinhard Pekrun, Anne C. Frenzel, Thomas Goetz, and Raymond P. Perry. The control-value theory of achievement emotions: An integrative approach to emotions in education. In Paul A. Schutz and Reinhard Pekrun, editors, *Emotion in Education*, pages 13–36. Academic Press, Burlington, January 2007.

[204] Reinhard Pekrun, Thomas Goetz, Anne C. Frenzel, Petra Barchfeld, and Raymond P. Perry. Measuring emotions in students' learning and performance: The achievement emotions questionnaire (aeQ). *Contemporary Educational Psychology*, 36(1):36–48, January 2011.

[205] Gabriela Trindade Perry, Ci Costa Kulpa, Eluza Pinheiro, and Marcelo Leandro Eichler. Lessons from an educational game usability evaluation. *International Journal of Interactive Mobile Technologies (iJIM)*, 6(2):23–28, April 2012.

[206] Rosalind W. Picard and Shaundra Bryant Daily. Evaluating affective interactions: Alternatives to asking what users feel. In *Proceedings of CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches (CHI)*, pages 2119–2122. ACM, 2005.

[207] David Pinelle, Nelson Wong, and Tadeusz Stach. Heuristic evaluation for games: Usability principles for video game design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1453–1462, Florence, Italy, April 2008. Association for Computing Machinery.

[208] Jan Plass, Bruce Homer, and Charles Kinzer. Playful Learning: An integrated design framework. Technical Report 02/2014, Games for Learning Institute, December 2014.

[209] Jan L. Plass, Suzy M. Letourneau, and Bruce D. Homer. Assessment mechanics: Design of learning activities that produce meaningful data. Learning Analytics and Educational Data Mining Workshop, April 2013. Published: New York University, CREATE Lab.

[210] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, May 2010.

[211] Kaśka Porayska-Pomsta, Manolis Mavrikis, Sidney D'Mello, Cristina Conati, and Ryan S. J. d. Baker. Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education*, 22(3):107–140, July 2013.

[212] Marc Prensky. Digital game-based learning. *Computers in Entertainment*, 1(1):21, October 2003.

[213] Rajagopal Raghunathan and Yaacov Trope. Walking the Tightrope between feeling ood and being accurate: Mood as a resource in processing persuasive messages. *Journal of Personality and Social Psychology*, 83(3):510–25, October 2002.

[214] O.M. Ramirez and Clarence J. Dockweiler. Mathematics anxiety: A systematic review. *Anxiety Stress & Coping: Advances in Test Anxiety Research*, 5:157–175, 1987.

[215] Rajib Rana, John Reilly, Raja Jurdak, Wen Hu, Xue Li, and Jeffrey Soar. Affect sensing on smartphone - Possibilities of understanding cognitive decline in aging population. *ArXiv*, 1407.5910 [cs], June 2015. arXiv: 1407.5910.

[216] Niklas Ravaja, Timo Saari, Mikko Salminen, Jari Laarni, and Kari Kallinen. Phasic emotional reactions to video game events: A psychophysiological investigation. *Media Psychology*, 8(4): 343–367, November 2006.

[217] C. Reuter, F. Mehm, Stefan Göbel, and R. Steinmetz. Evaluation of adaptive serious games using playtraces and aggregated play data. In *Proceedings of the 7th European Conference on Games Based Learning (ECGBL)*, volume 2, pages 504–511, Porto, Portugal, January 2013. Academic Conferences Ltd.

[218] Verónica Rivera-Pelayo, Valentin Zacharias, Lars Müller, and Simone Braun. A framework for applying quantified self approaches to support reflective learning. *Mobile Learning*, 2012:123–131, 2012.

[219] Ma. Mercedes T. Rodrigo, Ryan S. J. d. Baker, Sidney D'Mello, Ma. Celeste T. Gonzalez, Maria C. V. Lagud, Sheryl A. L. Lim, Alexis F. Macapanpan, Sheila A. M. S. Pascua, Jerry Q. Santillano, Jessica O. Sugay, Sinath Tep, and Norma J. B. Viehland. Comparing learners' affect while using an Intelligent Tutoring System and a simulation problem solving game. In Beverley P. Woolf, Esma Ameur, Roger Nkambou, and Susanne Lajoie, editors, *Intelligent Tutoring Systems*, pages 40–49. Springer, Berlin, Heidelberg, 2008.

[220] Elizabeth Rowe, Jodi Asbell-Clarke, and Ryan S. Baker. Serious Games Analytics to measure implicit science learning. In Christian Sebastian Loh, Yanyan Sheng, and Dirk Ifenthaler, editors, *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*, pages 343–360. Springer International Publishing, Cham, 2015.

[221] Jonathan P. Rowe, Bradford W. Mott, and James C. Lester. Opportunities and challenges in generalizable sensor-based affect recognition for learning. In *Proceedings of the Workshops at the 17th International Conference on Artificial Intelligence in Education (AIED)*, volume 6, pages 24–30, Raleigh, NC, 2015. Springer.

[222] Jennifer Sabourin, Jonathan P. Rowe, Bradford W. Mott, and James C. Lester. When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, pages 534–536. Springer, Springer, 2011.

[223] Jennifer L. Sabourin and James C. Lester. Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing*, 5(1):45–56, January 2014.

[224] Sergio Salmeron-Majadas, Olga C. Santos, and Jesus G. Boticario. Affective state detection in educational systems through mining multimodal data sources. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM)*, pages 348–349. ERIC, 2013.

[225] Maria Ofelia Z. San Pedro, Ryan S. J. d. Baker, and Ma. Mercedes T. Rodrigo. Carelessness and affect in an intelligent tutoring system for mathematics. *International Journal of Artificial Intelligence in Education (IJAIED)*, 24(2):189–210, June 2014.

[226] Jesse Schell. *The art of game design: A book of lenses*. CRC Press, Boston, MA, August 2008.

[227] Stefan Scherer, Nadir Weibel, Louis-Philippe Morency, and Sharon Oviatt. Multimodal prediction of expertise and leadership in learning groups. In *Proceedings of the 1st International Workshop on Multimodal Learning Analytics (MLA)*, 1, pages 1–8, Santa Monica, California, October 2012. Association for Computing Machinery.

[228] Magy Seif El-Nasr, Heather Desurvire, Lennart Nacke, Anders Drachen, Licia Calvi, Katherine Isbister, and Regina Bernhaupt. Game user research. In *Proceedings of the ACM CHI Annual Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, pages 2679–2682, Austin, Texas, USA, May 2012. Association for Computing Machinery.

[229] Ángel Serrano-Laguna, Eugenio J. Marchiori, Ángel del Blanco, Javier Torrente, and Baltasar Fernández-Manjón. A framework to improve evaluation in educational games. In *Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON)*, 148, pages 1–8, Marrakech, Morocco, April 2012. IEEE.

[230] Ángel Serrano-Laguna, Javier Torrente, Pablo Moreno-Ger, and Baltasar Fernández-Manjón. Application of learning analytics in educational videogames. *Entertainment Computing*, 5(4):313–322, December 2014.

[231] Noor Shaker and Mohammad Shaker. Towards understanding the nonverbal signatures of engagement in Super Mario Bros. In *Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization (UMAP)*, pages 423–434, Cham, 2014. Springer International Publishing.

[232] Noor Shaker, Stylianos Asteriadis, Georgios N. Yannakakis, and Kostas Karpouzis. Fusing visual and behavioral cues for modeling user experience in games. *IEEE Transactions on Cybernetics*, 43 (6):1519–1531, December 2013.

[233] David H. Shanabrook, Ivon Arroyo, and Beverly Park Woolf. Using touch as a predictor of effort: What the iPad can tell us about user affective state. In Judith Masthoff, Bamshad Mobasher, Michel C. Desmarais, and Roger Nkambou, editors, *User Modeling, Adaptation, and Personalization*, pages 322–327. Springer, Berlin, Heidelberg, 2012.

[234] Liping Shen, Enrique Leon, Victor Callaghan, and Ruimin Shen. Exploratory research on an affective e-learning model. In *Proceedings of the 2007 Workshop on Blended Learning (WBL)*, pages 267–278, Edinburgh, United Kingdom, January 2007. Academic Press.

[235] Kent Shi and Kamal Ali. GetJar Mobile application recommendations with very sparse datasets. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 204–212, Beijing, China, August 2012. Association for Computing Machinery.

[236] Ben Shneiderman. Designing for fun: How can we design user interfaces to be more fun? *Interactions*, 11(5):48–50, September 2004.

[237] Laila Shoukry and Stefan Göbel. Storyplay Multimodal: A research tool for the multimodal evaluation of serious games. In *Proceedings of the 11th European Conference on Games Based Learning (ECGBL)*, pages 577–584. Academic Conferences and Publishing International Limited, October 2017.

[238] Laila Shoukry and Stefan Göbel. Reasons and Responses: A multimodal serious games evaluation framework. *IEEE Transactions on Emerging Topics in Computing*, 8(1):245–255, January 2017.

[239] Laila Shoukry and Stefan Göbel. Realizing a mobile multimodal platform for serious games analytics. *International Journal of Serious Games (IJSG)*, 6(4):19–48, November 2019.

[240] Laila Shoukry, Stefan Göbel, and Ralf Steinmetz. Learning analytics and serious games: Trends and considerations. In *Proceedings of the 2014 ACM International Workshop on Serious Games (SeriousGames)*, pages 21–26, Orlando, Florida, USA, November 2014. Association for Computing Machinery.

[241] Laila Shoukry, Stefan Göbel, and Ralf Steinmetz. Towards Mobile Multimodal Learning Analytics. In *Proceedings of the Learning Analytics for and in Serious Games Workshop at the European Conference of Technology Enhanced Learning (ECTEL)*, page 16, Darmstadt, Germany, 2014.

[242] Laila Shoukry, Christian Reuter, and Florian Mehm. StoryTec and StoryPlay as tools for adaptive game-based learning research. In Stefan Göbel and Josef Wiemeyer, editors, *Games for Training, Education, Health and Sports*, pages 195–198. Springer International Publishing, Cham, 2014.

[243] Laila Shoukry, Christian Sturm, Galal H. Galal-Edeen, and Stefan Göbel. Conducting evaluation studies of mobile games with preschoolers. In *Proceedings of DeLFI Workshops 2014 co-located with 12th e-Learning Conference of the German Computer Society (DeLFI)*, volume 1227, pages 262–269, Freiburg, Germany, 2014. CEUR-WS.org.

[244] Laila Shoukry, Polona Caserman, Stefan Göbel, and Ralf Steinmetz. Blitzmerker: Learning idioms with a mobile game. In Stefan Göbel, Minhua Ma, Jannicke Baalsrud Hauge, Manuel Fradinho Oliveira, Josef Wiemeyer, and Viktor Wendel, editors, *Serious Games*, pages 77–88. Springer International Publishing, Cham, 2015.

[245] Laila Shoukry, Christian Sturm, and Galal H. Galal-Edeen. Pre-MEGa: A proposed framework for the design and evaluation of preschoolers' mobile educational games. In Tarek Sobh and Khaled Elleithy, editors, *Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering*, pages 385–390. Springer International Publishing, Cham, 2015.

[246] Laila Shoukry, Johannes Konert, and Stefan Göbel. The evaluation of learner experience in serious games. In Imed Bouchrika, Nouzha Harrati, and Phu Vu, editors, *Learner Experience and Usability in Online Education*, pages 122–148. IGI Global, January 2017.

[247] Valerie J. Shute. Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, 55(2):503–524, 2011.

[248] Valerie J. Shute, Matthew Ventura, Malcolm Bauer, and Zapata-Rivera Diego. Melding the power of serious games and embedded assessment to monitor and foster learning: flow and grow. In

Ute Ritterfeld, Michæl Cody, and Peter Vorderer, editors, *Serious Games: Mechanisms and Effects*, pages 295–321. Routledge, New York, August 2009.

[249] George Siemens and Ryan S. J. d. Baker. Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK)*, pages 252–254, Vancouver, British Columbia, Canada, April 2012. Association for Computing Machinery.

[250] George Siemens and Phil Long. Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46(5):30, 2011.

[251] Anna Skinner, Clementina Russo, Lisa Baraniecki, and Molly Maloof. Ubiquitous Augmented Cognition. In DylanD. Schmorrow and CaliM. Fidopiastis, editors, *Foundations of Augmented Cognition. Advancing Human Performance and Decision-Making through Adaptive Systems*, volume 8534, pages 67–77. Springer International Publishing, 2014.

[252] Debra J. Slone. A Methodology for measuring usability evaluation skills using the constructivist theory and the Second Life virtual world. *Journal of Usability Studies*, 4(4):178–188, August 2009.

[253] Jiqiang Song, Eugene Y. Tang, and Leibo Liu. User behavior pattern analysis and prediction based on mobile phone sensors. In Chen Ding, Zhiyuan Shao, and Ran Zheng, editors, *Network and Parallel Computing*, pages 177–189. Springer, Berlin, Heidelberg, 2010.

[254] Minzhu Song and Sujing Zhang. EFM: A model for educational game design. In Zhigeng Pan, Xiaopeng Zhang, Abdennour El Rhalibi, Woontack Woo, and Yi Li, editors, *Technologies for E-Learning and Digital Entertainment*, pages 509–517. Springer, Berlin, Heidelberg, 2008.

[255] Jacopo Staiano, María Menéndez, Alberto Battocchi, Antonella De Angeli, and Nicu Sebe. UX_Mate: from facial expressions to UX evaluation. In *Proceedings of the Designing Interactive Systems Conference (DIS)*, pages 741–750, Newcastle Upon Tyne, United Kingdom, June 2012. Association for Computing Machinery.

[256] Christina Steiner, P. Hollins, Eric Kluijfhout, Mihai Dascalu, Alexander Nussbaumer, Dietrich Albert, and Wim Westera. Evaluation of serious games: A holistic approach. In *Proceedings of the 8th International Conference of Education Research and Innovation (ICERI)*, pages 4334–4342, Sevilla, Spain, November 2015. IATED.

[257] Christina M. Steiner, Michael Kickmeier-Rust, Elke Mattheiss, Stefan Göbel, and Dietrich Albert. Balancing on a high wire: Adaptivity key factor of future learning games. In Dr Michæl D. Kickmeier-Rust and Dr Dietrich Albert, editors, *An Alien's Guide to Multi-Adaptive Educational Computer Games*, pages 43–88. Informing Science Press, February 2012.

[258] Ralf Steinmetz. Synchronization properties in multimedia systems. *IEEE Journal on Selected Areas in Communications*, 8(3):401–412, 1990.

[259] Sophie Stellmach, Lennart Nacke, and Raimund Dachselt. Advanced gaze visualizations for three-dimensional virtual environments. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA)*, pages 109–112, Austin, Texas, March 2010. Association for Computing Machinery.

[260] Veronica Sundstedt. *Gazing at Games: An introduction to eye tracking control*. Morgan & Claypool Publishers, San Rafæl, CA, March 2012.

[261] Sakari Tamminen, Antti Oulasvirta, Kalle Toiskallio, and Anu Kankainen. Understanding Mobile Contexts. *Personal and Ubiquitous Computing*, 8(2):135–143, May 2004.

[262] Chek Tien Tan, Sander Bakkes, and Yusuf Pisan. Inferring player experiences using facial expressions analysis. In *Proceedings of the 2014 Conference on Interactive Entertainment (IE)*, 7, pages 1–8, Newcastle, NSW, Australia, December 2014. Association for Computing Machinery.

[263] Chek Tien Tan, Pejman Mirza-Babæi, Veronica Zammitto, Alessandro Canossa, Genevieve Conley, and Guenter Wallner. Tool Design Jam: Designing tools for Games User Research. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY)*, pages 827–831, London, United Kingdom, October 2015. Association for Computing Machinery.

[264] Kikumi K. Tatsuoka. Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In Paul D. Nichols, Susan F. Chipman, and Robert L. Brennan, editors, *Cognitively Diagnostic Assessment*, pages 327–359. Routledge, Hillsdale, N.J, March 1995.

[265] Edison Thomaz, Aman Parnami, Jonathan Bidwell, Irfan Essa, and Gregory D. Abowd. Technological approaches for addressing privacy concerns when recognizing eating behaviors with wearable cameras. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 739–748, Zurich, Switzerland, September 2013. Association for Computing Machinery.

[266] Kate Thompson. Using micro-patterns of speech to predict the correctness of answers to Mathematics problems: An exercise in multimodal learning analytics. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI)*, pages 591–598, Sydney, Australia, December 2013. Association for Computing Machinery.

[267] Marko Tkalcic, Ante Odic, Andrej Kosir, and Jurij Tasic. Affective labeling in a content-based recommender system for images. *IEEE Transactions on Multimedia*, 15(2):391–400, February 2013.

[268] John Traxler. Defining, discussing and evaluating mobile learning: The moving finger writes and having writ . . . . *The International Review of Research in Open and Distributed Learning*, 8(2), June 2007.

[269] Jan-Paul van Staalduinen and Sara de Freitas. A game-based learning framework: Linking game design and learning outcomes. In Myint Swe Khine, editor, *Learning to Play: Exploring the Future of Education with Video Games*, pages 29–54. Peter Lang, New York, NY, 2011.

[270] Giasemi Vavoula and Mike Sharples. Meeting the challenges in evaluating mobile learning: A 3-level evaluation framework. *International Journal of Mobile and Blended Learning*, 1(2):54–75, 2009.

[271] George Veletsianos. Contextually relevant pedagogical agents: Visual appearance, stereotypes, and first impressions and their impact on learning. *Computers & Education*, 55(2):576–585, September 2010.

[272] Terry P. Vendlinski, Girlie C. Delacruz, Rebecca E. Buschang, Gregory K. W. K. Chung, and Eva L. Baker. Developing high-quality assessments that align with instructional video games. Technical Report CRESST Report 774, ERIC Clearinghouse, 2010. Google-Books-ID: SZT2vgEACAAJ.

[273] Adrian Vosskühler. OGAMA description (for version 2.5) : A software to record, analyze and visualize gaze and mouse movements in screen-based environments. Technical report, Freie Universität Berlin, 2010. event-place: Berlin, Germany.

[274] Van Den W. M. Hoogen, W. A. IJsselsteijn, and De Y. A.W. Kort. Towards real-time behavioral indicators of player experiences : pressure patterns and postural responses. In *Proceedings of the 6th International Conference on Methods and Techniques in Behavioral Research*, pages 100–101, Maastricht, 2008. Noldus.

[275] G. Wallner and S. Kriglstein. Visualization-based analysis of gameplay data : A review of literature. *Entertainment Computing*, 4(3):143–155, August 2013.

[276] Althea Vail Wallop. Learning to predict disengagement in educational iPad application. Technical report, Althea Vail Wallop CS 229 Final Project Report, 2010.

[277] Ben George Weber, Michæl John, Michæl Mateas, and Arnav Jhala. Modeling player retention in Madden NFL 11. In *Proceedings of the twenty-third Innovative Applications of Artificial Intelligence Conference (IAAI)*, San Francisco, CA, 2011. AAAI.

[278] Nadir Weibel, Adam Fouse, Colleen Emmenegger, Sara Kimmich, and Edwin Hutchins. Let's look at the cockpit: Exploring mobile eye-tracking for observational research on the flight deck. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA)*, pages 107–114, Santa Barbara, California, March 2012. Association for Computing Machinery.

[279] Benjamin Weiss. Analyzing user behavior for quality assessment. In *Proceedings of the International Conference Interfaces and Human Computer Interaction (IADIS )*, pages 223–226. IHE, 2012.

[280] Grant Wiggins and Jay McTighe. *The understanding by design guide to creating high-quality units*. ASCD, Alexandria, Va, January 2011.

[281] Brian M. Winn. The design, play, and experience framework. In Richard E. Ferdig, editor, *Handbook of Research on Effective Electronic Gaming in Education Set of 3*, pages 1010–1024. Information Science Reference, Hershey, PA, August 2008.

[282] Marcelo Worsley. Multimodal Learning Analytics as a tool for bridging learning theory and complex learning behaviors. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge (MLA)*, 2, pages 1–4, Istanbul, Turkey, November 2014. Association for Computing Machinery.

[283] Marcelo Worsley and Paulo Blikstein. Deciphering the practices and affordances of different reasoning strategies through multimodal learning analytics. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge (MLA)*, pages 21–27, Istanbul, Turkey, November 2014. Association for Computing Machinery.

[284] Jane Yin-Kim Yau and Mike Joy. An adaptive context-aware mobile learning framework based on the usability perspective. *International Journal of Mobile Learning and Organisation*, 4(4): 378–390, December 2010.

[285] Chungkuk Yoo, Jæmyung Shin, Inseok Hwang, and Junehwa Song. FaceLog: Capturing user's everyday face using mobile devices. In *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 163–166, Zurich, Switzerland, September 2013. Association for Computing Machinery.

[286] Bieke Zaman and Vero Vanden Abeele. Towards a likeability framework that meets child-computer interaction & communication sciences. In *Proceedings of the 6th International Conference on Interaction Design and Children (IDC)*, 1, pages 1–8, Aalborg, Denmark, June 2007. Association for Computing Machinery.

[287] Veronica Zammitto, Pejman Mirza-Babæi, Ian Livingston, Marina Kobayashi, and Lennart E. Nacke. Player experience: mixed methods and reporting results. In *Proceedings of the ACM CHI Annual Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA)*, pages 147–150, Toronto, Ontario, Canada, April 2014. Association for Computing Machinery.

[288] Juan-Diego Zapata-Rivera and Jim E. Greer. Interacting with inspectable Bayesian student models. *International Journal of Artificial Intelligence in Education (IJAIED)*, 14(2):127–163, April 2004.

[289] Panagiotis Zervas, Sergio Eduardo Gomez Ardila, Ramon Fabregat, and Demetrios G. Sampson. Tools for context-aware learning design and mobile delivery. In *Proceedings of the 2011 IEEE 11th International Conference on Advanced Learning Technologies (ICALT)*, pages 534–535, USA, July 2011. IEEE Computer Society.

[290] Patrick H. Zimmerman, J. Elizabeth Bolhuis, Albert Willemsen, Erik S. Meyer, and Lucas P. J. J. Noldus. The Observer XT: A tool for the integration and synchronization of multimodal signals. *Behavior Research Methods*, 41(3):731–735, August 2009.

## List of Figures

## List of Tables

## Abbreviations

| | |
|---|---|
| **KOM** | Multimedia Communications Lab at the Technical University of Darmstadt |
| **NGLOB** | Narrative Game-based Learning Object |
| **MM** | Multimodal |
| **LA** | Learning Analytics |
| **MMLA** | Multimodal Learning Analytics |
| **EDM** | Educational Data Mining |
| **CbKST** | Competence-based Knowledge Space Theory |
| **OLM** | Open Learner Model |
| **ECD** | Evidence-centered Design |
| **SAE** | Skill Assessment Engine |
| **MAE** | Motivation Assessment Engine |
| **SC** | Skin Conductance |
| **HRV** | Heart Rate Variability |
| **HR** | Heart Rate |
| **GSR** | Galvanic Skin Response |
| **EMG** | Electromyography |
| **fEMG** | Facial Electromyography |
| **EEG** | Electroencephalogram |
| **EKG** | Electrocardiogram |
| **CSCL** | Computer Supported Collaborative Learning |
| **MMLA** | Multimodal Learning Analytics |
| **SCROLL** | System for Capturing and Reminding of Ubiqitous Learning Log |
| **PC** | Personal Computer |
| **GT** | Ground Truth |
| **BROMP** | Baker Rodrigo Ocumpaugh Monitoring Protocol |
| **BCI** | Brain-Computer Interface |
| **EDA** | Event-Driven Architecture |
| **AAM** | Active Appearance Model |

| | |
|---|---|
| **OS** | Operating System |
| **CPU** | Central Processing Unit |
| **QDAS** | Qualitative Data Analysis Software |
| **TCP** | Transmission Control Protocol |
| **SDK** | Software Development Kit |
| **ICML** | INSCAPE Markup Language |
| **GUI** | Graphical User Interface |
| **WPF** | Windows Presentation Foundation |
| **UX** | User eXperience |
| **HTML5** | Hypertext Markup Language version 5 |
| **LAIF** | A Logging and Interaction Framework for Gaze-Based Interfaces |
| **GLEANER** | Game Learning Analytics for Educational Research |
| **TRUE** | Tracking Real-time User Experience |
| **UoLmP** | Units of Learning mobile Player |
| **ColAT** | Collaboration Analysis Tool |
| **LeGUC** | Learning-Gaming-Using-Context Model |
| **CPV** | Clicks per Visit |
| **TPV** | Time per Visit |
| **LMS** | Learning Management System |
| **DRS** | Digital Replay System |
| **FBGM** | Feedback-based Gameplay Metrics |

# Appendix

## A List of Papers Arising from this Thesis

Laila Shoukry and Stefan Göbel. Reasons and Responses: A multimodal serious games evaluation framework. *IEEE Transactions on Emerging Topics in Computing*, 8(1):245–255, January 2017

Laila Shoukry and Stefan Göbel. Realizing a mobile multimodal platform for serious games analytics. *International Journal of Serious Games (IJSG)*, 6(4):19–48, November 2019

Laila Shoukry, Johannes Konert, and Stefan Göbel. The evaluation of learner experience in serious games. In Imed Bouchrika, Nouzha Harrati, and Phu Vu, editors, *Learner Experience and Usability in Online Education*, pages 122–148. IGI Global, January 2017

Laila Shoukry and Stefan Göbel. Storyplay Multimodal: A research tool for the multimodal evaluation of serious games. In *Proceedings of the 11th European Conference on Games Based Learning (ECGBL)*, pages 577–584. Academic Conferences and Publishing International Limited, October 2017

Laila Shoukry, Stefan Göbel, and Ralf Steinmetz. Learning analytics and serious games: Trends and considerations. In *Proceedings of the 2014 ACM International Workshop on Serious Games (SeriousGames)*, pages 21–26, Orlando, Florida, USA, November 2014. Association for Computing Machinery

Laila Shoukry, Christian Sturm, Galal H. Galal-Edeen, and Stefan Göbel. Conducting evaluation studies of mobile games with preschoolers. In *Proceedings of DeLFI Workshops 2014 co-located with 12th e-Learning Conference of the German Computer Society (DeLFI)*, volume 1227, pages 262–269, Freiburg, Germany, 2014. CEUR-WS.org

Laila Shoukry, Stefan Göbel, and Ralf Steinmetz. Towards Mobile Multimodal Learning Analytics. In *Proceedings of the Learning Analytics for and in Serious Games Workshop at the European Conference of Technology Enhanced Learning (ECTEL)*, page 16, Darmstadt, Germany, 2014

Laila Shoukry, Christian Reuter, and Florian Mehm. StoryTec and StoryPlay as tools for adaptive game-based learning research. In Stefan Göbel and Josef Wiemeyer, editors, *Games for Training, Education, Health and Sports*, pages 195–198. Springer International Publishing, Cham, 2014

Laila Shoukry, Christian Sturm, and Galal H. Galal-Edeen. Pre-MEGa: A proposed framework for the design and evaluation of preschoolers' mobile educational games. In Tarek Sobh and Khaled Elleithy, editors, *Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering*, pages 385–390. Springer International Publishing, Cham, 2015

Laila Shoukry, Polona Caserman, Stefan Göbel, and Ralf Steinmetz. Blitzmerker: Learning idioms with a mobile game. In Stefan Göbel, Minhua Ma, Jannicke Baalsrud Hauge, Manuel Fradinho Oliveira, Josef Wiemeyer, and Viktor Wendel, editors, *Serious Games*, pages 77–88. Springer International Publishing, Cham, 2015

## B Online Evaluation Forms with Instructions and Responses

### B.1 Online Evaluation Form for the Capturer App

The following is the online form sent to participants to evaluate the process of playtesting games using the capturer app of StoryPlay Multimodal. It is exported here from Google Forms.

# Playtesting your Game

Thank you very much for taking part in this testing. By doing so you are helping advance research of educational games evaluation. Also, it will help you learn more about the process of game playtesting.

شكرا لكم لمشاركتكم في هذا التقييم والمساعدة في ابحاث الالعاب الالكترونية التعليمية.

If you have any problems in installation or have any questions or want to be updated with new games, please join this group:

لو فيه أي مشاكل قابلتكوا او عندكوا اي استفسارات او عايزين

ياريت تخشوا على الجروب دة

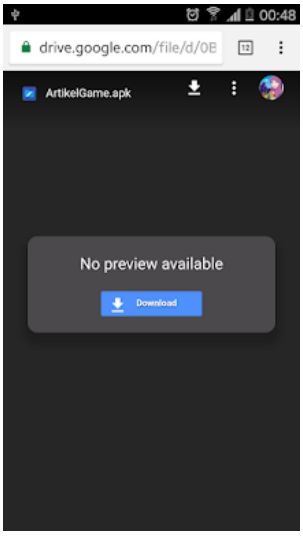
https://chat.whatsapp.com/FQRZEhReX0V4TrdlKN0DsS

* Required

Instructions for running the game


1. Please write your smartphone or tablet model and android version (the one you will use for evaluation) *

Download the apk on your mobile or transfer it from your computer to your mobile using usb connection.



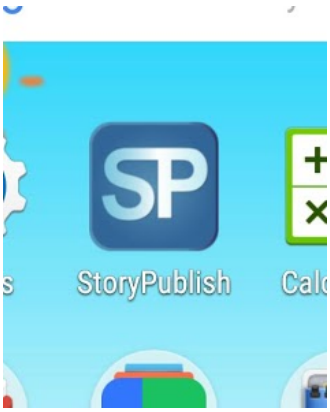

2. When you try installing it, it will ask you to allow installing an app from outside google play. So you check that you allow it as in the picture below: (allow unknown sources)

لما تنزلوها وتيجو تعملوا install حيقوللكوا تأكدوا انكوا موافقين تحطوا لعبة مش جاية من google play . فعشان تنزلوا اللعبة لازم تضغطوا انكوا موافقين زي الصورة اللي تحت:


3. Running the App:

تشغيل البرنامج

After installing, the app will look like the below picture: You can put it in an accessible place so you can open easily. بعد تنزيل اللعبة حيبقى دة شكلها زي الصورة: ياريت تحطوها في مكان واضح .عشان تلاقوها بعد كدة.




2. Make sure all requested permissions are given for this app by going to Settings --> Apps --> StoryPublish --> permissions --> and checking all permissions (camera, location, etc.). Were they unchecked on your device or already checked? *

   *Mark only one oval.*

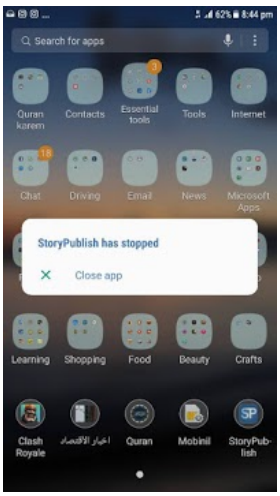
   - they were unchecked
   - they were checked
   - I couldn't do this part

When you run the app, this screen will show up: The camera is checked which means the face of the tester will be video recorded. If you don't want that, you uncheck it. Also the location is checked which means GPS data will be collected. Other sensors can be deactivated using the third button too. Please create a new profile for every tester who will play the game by clicking on New Profile and filling out the data. Then every time you can choose his name from the list of profiles before playing. لما تفتحوها تظهر الشاشة دي. علامة الكاميرا يعني حتصور وش الطفل أو التستر وهو بيلعب. لو مش عايزينها بتلغوها (بس حنحتاجها في الخطوات القادمة) ونفس الموضوع بالنسبة لباقي ال sensors. ياريت تعملوا بروفايل جديد لكل طفل حيستخدم البرنامج وبعد كدة حتختاروا اسمه كل مايستخدمها.

4. filling out profile data: Please fill out the profile data for each tester using the form below. Then click on confirm and start game.

ياريت تملوا بيانات كل طفل في الشاشة دي وبعدين تدوسوا confirm و Start Game .


If the app closes and this error appears, then please run it again and disable the camera this time (uncheck the camera icon), then it will run. لو البرنامج قفل وظهر الشكل دة ياريت تشغلوه تاني وتلغوا الكاميرا المرة دي حيشتغل ان شاء الله.




3. Did the app crash when the camera option was checked?

*Mark only one oval.*

- yes
- No
- I never checked the camera option

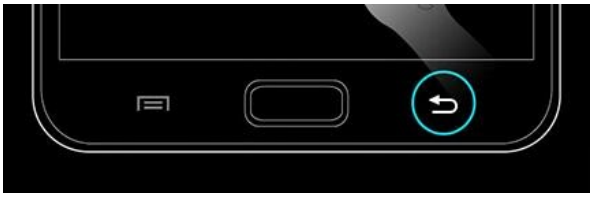
4. Now Play the game or let your tester play it! If you are the one who is testing, please think-aloud while testing to show us what you like/dislike/find boring/confusing/exciting about the game. Please write here your name and the name of the games you will test.

_____

## Finishing the Game

إنهاء اللعبة

When the tester finishes playing the game, please press the back button of your mobile. It looks like this: بعد ما الطفل يخلص لعب ياريت تدوس على زرار العودة بتاع الموبايل ودة شكله:

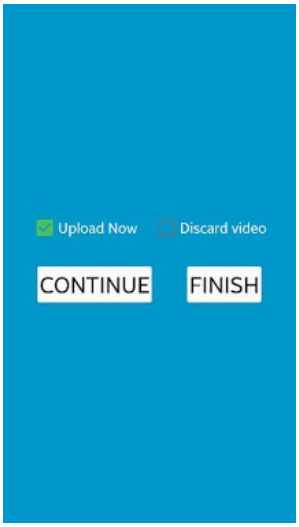


This rating screen will appear: Here you choose the rating from the point of view of the tester or explain and ask them to do it themselves. Dislike/Bored/Confused/Excited/Loved it. حتظهر الشاشة دي لتقييم اللعبة. ياريت تختاروا التقييم من وجهة نظر اشخص اللي لعبها أو اشرحوهاله وخلوه هو يختار: مش عاجباه أو مملة أو مش مفهومة أو مسلية أو رائعة

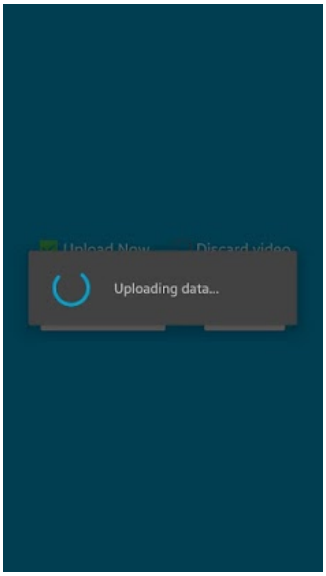

## Uploading Data

إرسال البيانات

After you choose you will get this screen. If you don't have internet you can choose to upload the data later by unchecking "upload now". If you don't want the recorded video to be sent you can check "discard video" but if it is not a problem for you please let it be uploaded. If you clicked the back button by mistake, then you can continue the game by clicking continue. To finish the game click Finish. بعد ماتختاروا حتظهرلكوا الشاشة دي. هنا لو مفيش نت دلوقت وعايزين تبعتوا الداتا بعدين ممكن تشيلوا العلامة من upload now. ولوعايزين تمسحوا الفيديو اللي اتسجل ومايتبعتش ممكن تختاروا Discard Video بس حيفيدنا جدا لو بعتوه. ولوكنتو دستو رجوع بالغلط وعايزين تكملوا اللعبة دوسو continue. دوسوا finish لانهاء البرنامج وارسال البيان.




While sending data, this screen will appear. This means you have done everything correctly, thank you very much! Please at this point to continue using your mobile while uploading use the Home button of your mobile and not the back button, as the back button will cancel uploading. The home screen will let you continue working normally and the uploading will be done in the background and then closes the app. اثناء ارسال البيانات بتظهر الشاشة دي. لو عايزين تكملوا شغل على الموبايل اللي بيبقى في نص الموبايل وتقدروا تشتغلوا عادي. لو وصلتوا لغاية هنا Home عشان حيلغي الارسال. ياريت تدوسوا على زرار back عادي ياريت ماتدوسوش على يبقى بجد شكرا ليكو وليكوا هدية عندي ان شاء الله.



5. Did you see the upload screen? *

*Mark only one oval.*

- Yes
- No

6. Please let someone take a picture of you while you are using the applications which shows you and the applications you are testing and upload these pictures here. I will check the data uploaded by your app, too. And you will need it for your next milestone. Thank you. *

Files submitted:

Thank you. Please answer the following Questions

7. How difficult was carrying out this playtesting session? *

*Mark only one oval.*

- Very Difficult
- Difficult
- Medium
- Easy

8. Did you encounter any bugs in StoryPublish? *

*Mark only one oval.*

- The Camera didn't work
- The uploading crashed
- no bugs
- Other: _____

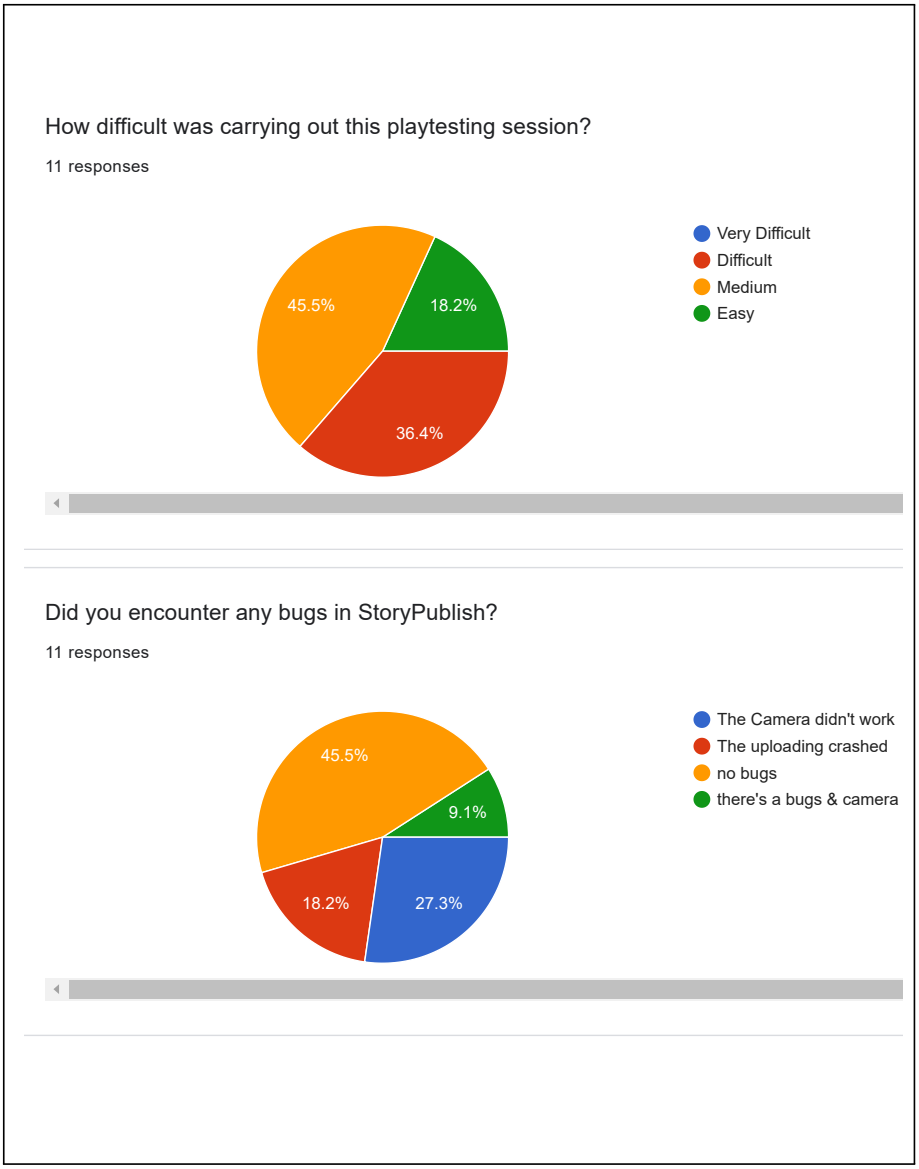
9. How easy and intuitive was it to use Storypublish? *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Very Easy | | | | | | Very Difficult |

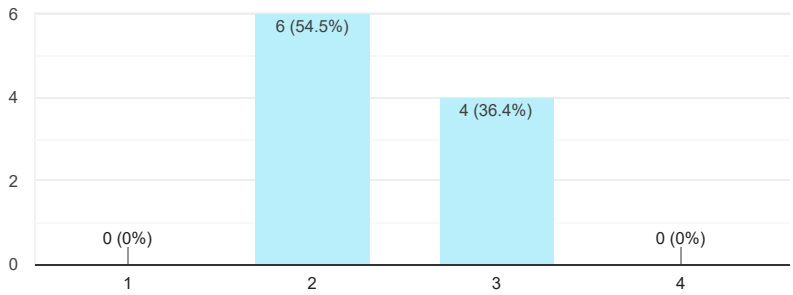
## B.2 Some Evaluation Form Results of the Capturer App

The following are some results from the evaluation form.



How easy and intuitive was it to use Storypublish?

11 responses




Do you think the data collected using Storypublish will be useful for evaluating the gar
Please explain.

11 responses

yes

yes it required some permissions which is useful like the camera to record while playing but it doesn't work with me and make profile to collect data about the user who play the game

maybe

yes

i think that but there was a big pro it's independ on one software

Ja,weil es einen Lernerfolg gibt und man diesen auch Vergleichen kann.Man sieht auch wie die Kinder waehrend des Spiel fuehlen und benahmen.Auch wie die betreuende Perdon die aufgenommen hat.Man kann auch einen Lernerfolg durch Lernprogramme nachweisen.Insbes durch dieses

## B.3 Online Evaluation Form for the Observer App

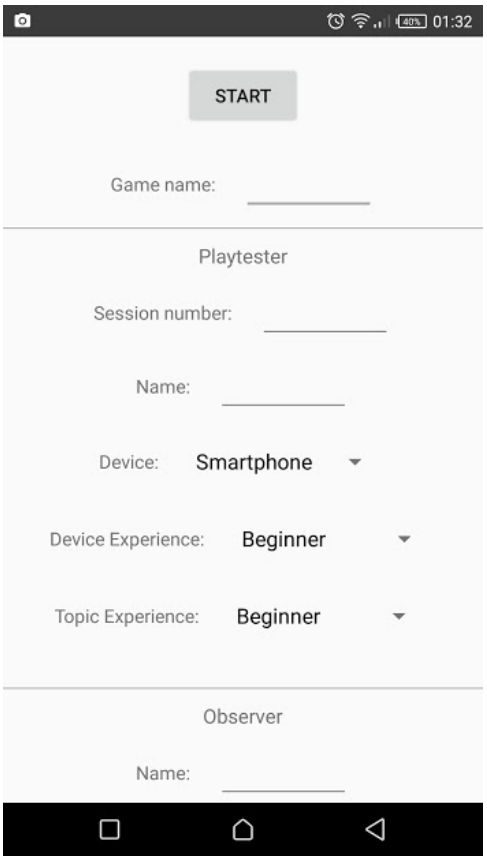
The following is the online form sent to participants to evaluate the process of annotating playtesting observations using the observer app of StoryPlay Multimodal. The form is exported here from Google Forms.

# Using the Observer app

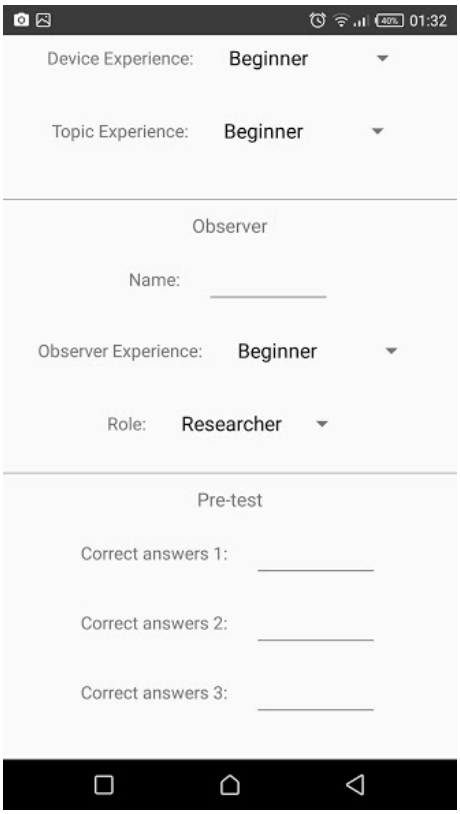
This app will be used for taking fast observation notes while someone is playing your game. First, please download and install the observer apk on your device. After installing, go to Settings, choose applications, choose observer, choose permissions and give it the permission it needs for storage. You will need another device where your game is already installed so that you use the current device for observations while the tester is playing your game. When you install it and run it, you will see the following screen:

* Required

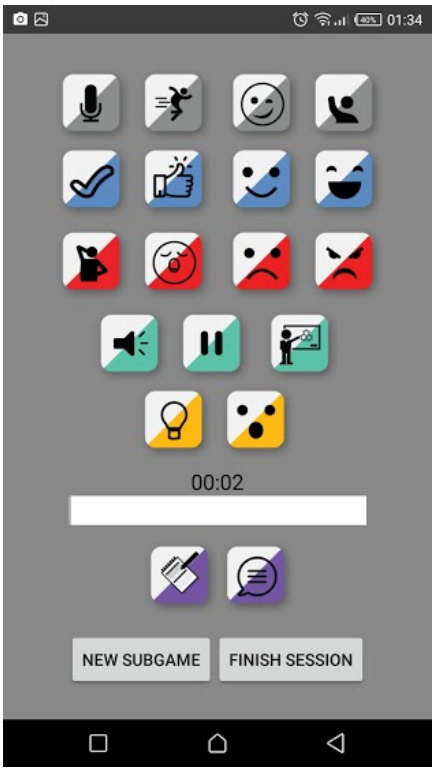
Please fill out your game name, session number (displayed at the start of the game) and other information about your tester and about you (the observer). During this time, also make sure to fill the profile information of the tester on your game app and be sure that the session number is the same in both apps.



You can leave the pretest spaces blank. Now both of you start your sessions at the same instance by clicking on start.



On the other mobile, the game will start. On your observer app, you will see the following screen with many icons, a textfield and some buttons. You click on the icons when you observe a certain behavior, context or reaction which happens during the testing. For example, the grey icons stand for the tester doing one of those behaviors: speaking, moving, tricking, asking for help. The blue icons stand for his positive reactions like being confident while answering a question, being proud of answering correctly, smiling or laughing. The red icons stand for negative reactions of the tester like being confused, bored, sad or angry. The green icons are for context events like noise, interruption, or you offering help to the tester and explaining something in the game. The yellow icons stand for nutral reactions like reflection and surprise. You can use the textfield for two things: taking a note about something like a bug for example and then clicking the left purple button, or writing something that the player said and then clicking the right purple button. The new subgame button can be used when a player goes to the main menu and choose a new subgame. When the player wishes to finish the session, please click on Finish session and your observed data will be automatically uploaded along with its timestamp in a .log file.




Thank you. Please answer the following Questions

1. How difficult was carrying out this playtesting session? *

*Mark only one oval.*

- Very Difficult
- Difficult
- Medium
- Easy

2. Was the testing too fast for taking notes? How many observations have you missed because of speed? *

*Mark only one oval.*

- many
- some
- few
- none

3. Did anything happen which you didn't find a button for? If yes, please specify

4. Did you start your sessions simultaneously? *

*Mark only one oval.*

- Yes exactly
- Yes but with some lag
- No, there was a big difference in time

5. Do you think this app is useful for recording observations during playtesting? What is the difference/advantage of it to taking notes using pen and paper? *

6. Do you have any recommendations for improving this app? *

7. Did you notice anything you need to improve in your game during this playtesting session? What is it? *

**Thank you for your efforts.**

8. Please take a photo of your testers while playing your game and let someone take a photo of you while using the observer app and upload these pictures here. I will check the data uploaded by your app, too. And you will need it for your next milestone. Thank you. *

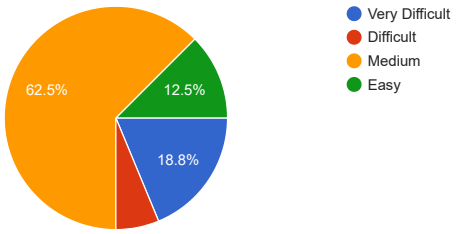
   Files submitted:

## B.4 Some Evaluation Form Results of the Observer App

The following are some results from the evaluation form.

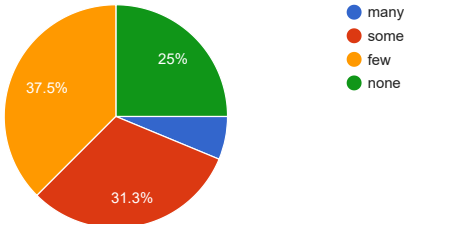
How difficult was carrying out this playtesting session?

16 responses



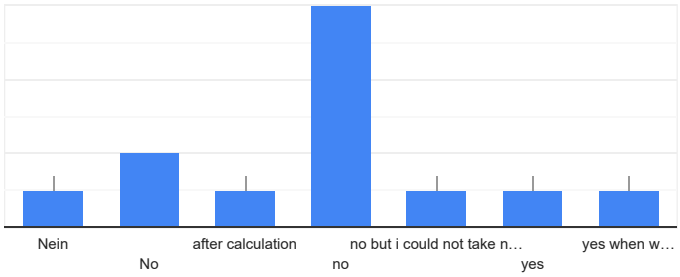

Was the testing too fast for taking notes? How many observations have you missed because of speed?

16 responses



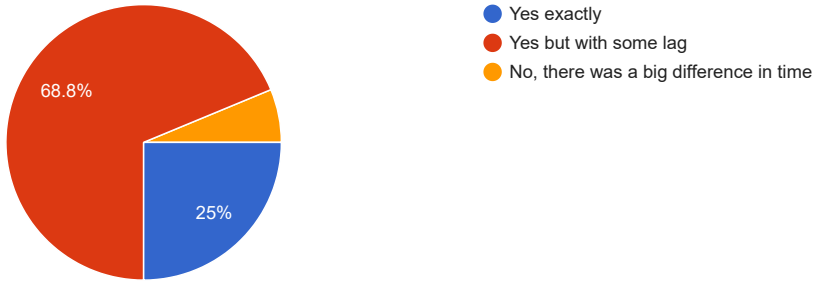

Did anything happen which you didn't find a button for? If yes, please specify

13 responses



Did you start your sessions simultaneously?

16 responses




Do you think this app is useful for recording observations during playtesting? What is the difference/advantage of it to taking notes using pen and paper?

16 responses

yes

No

it was useful , but taking notes was difficlut

no not useful, the adv of taking notes to prevent forget any comment from the player

yes the difference is the time of taking note may take more time & i miss some of notes

maybe

paper can lose

yes it is more helpful than pen and paper and more funny and easy

no

## B.5 Online Evaluation Form for Game Testing

The following are the online forms sent to participants to evaluate the game of another student which they tested using the capturer app, so that other students can improve their games based on the feedback given by testers. The form is exported here from Google Forms.

# Evaluating your Friends' Game

Please answer these questions to tell your opinion about the game you have tested.

* Required

1. Email address *

2. Please write your name here: *

3. Please write the name of the game you have tested. *

Please answer the following Questions:

4. How difficult was using the game? *

*Mark only one oval.*

- Very Difficult
- Difficult
- Medium
- Easy

5. Did you encounter any bugs in the game? *

*Mark only one oval.*

- many
- some
- few
- none

6. Please rate your enjoyment of the game *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Didn't enjoy at all | | | | | | Enjoyed a lot |

7. Which part did you like most in the game? *

8. Do you have any suggestions for improving the game? *

Thank you for your efforts. If you have any further comments, please add them here:

## B.6 Online Evaluation Form for ArtikelGame Testing

The following are the online forms sent to participants to evaluate the Artikelgame which was specifically created to collect data for the purpose of this research. The form is exported here from Google Forms.

# Artikel Game Testing

Thank you very much for taking part in this testing. By doing so you are helping us advance research of educational games evaluation.

شكرا لكم لمشاركتكم في هذا التقييم ومساعدتنا في ابحاثنا في مجال الالعاب الالكترونية التعليمية.

If you have any problems in installation or have any questions or want to be updated with new games, please join this group:

https://chat.whatsapp.com/77neVBKeABCC8BynP5RPCA

لو فيه أي مشاكل قابلتكوا او عندكوا اي استفسارات او عايزين

تتابعوا لعبنا الجديدة ياريت تخشوا على الجروب دة

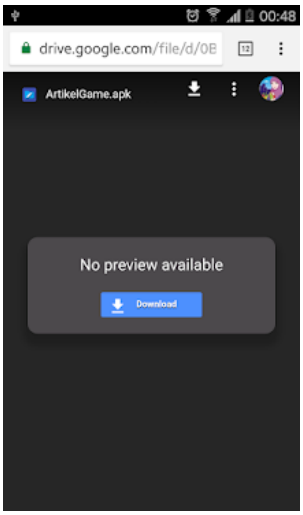
https://chat.whatsapp.com/77neVBKeABCC8BynP5RPCA

Instructions for running the game


1. Please dowload the game using this link:

برجاء تنزيل اللعبة من هذا الرابط:

https://drive.google.com/file/d/0B7gaXCr-uaE2Ny1HWUJPSmlxUzg/view?usp=drivesdk

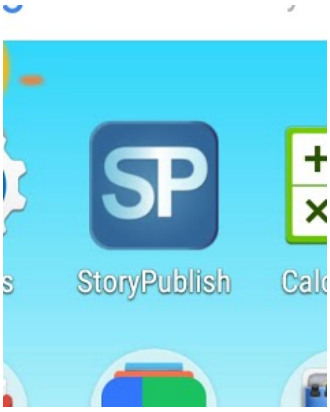


لما تنزلوها وتيجو تعملوا install حيقوللكوا تأكدوا انكوا موافقين تحطوا لعبة مش جاية من google play . فعشان تنزلوا اللعبة لازم تضغطوا انكوا موافقين زي الصورة اللي تحت:


تشغيل البرنامج

3. Running the App:

After installing, the app will look like the below picture: You can put it in an accessible place so you can open easily. Please run the app. بعد تنزيل اللعبة حيبقى دة شكلها زي الصورة: ياريت تحطوها في مكان واضح عشان تلاقوها بعد كدة. وبعدين شغلوها.

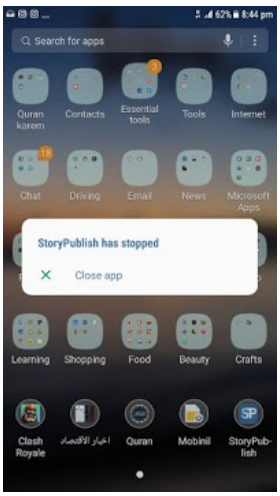


When you open it, this screen will show up: The camera is checked which means the face of the child will be video recorded. If you don't want that, you uncheck it. Also the location is checked which means GPS data will be collected. Other sensors can be deactivated using the third button too. Please create a new profile for every child who will play the game by clicking on New Profile and filling out the data. Then every time you can choose his name from the list of profiles before playing. (لما تفتحوها تظهر الشاشة دي. علامة الكاميرا يعني حتصور وش الطفل وهو بيلعب. لو مش عايزينها بتلغوها (بس محتاجاها في البحث
ونفس الموضوع بالنسبة لباقي ال sensors. ياريت تعملوا بروفايل جديد لكل طفل حيستخدم البرنامج وبعد كدة حتختاروا اسمه كل مايستخدمها.


4. filling out profile data: Please fill out the profile data for each child using the form below. Then click on confirm and start game.

ياريت تملوا بيانات كل طفل في الشاشة دي وبعدين تدوسوا confirm و Start Game .

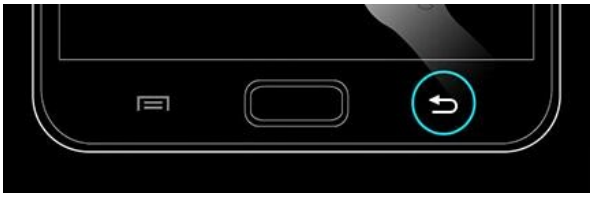
If the app closes and this error appears, then please run it again and disable the camera this time (uncheck the camera icon), then it will run. لو البرنامج قفل وظهر الشكل دة ياريت تشغلوه تاني وتلغوا الكاميرا المرة دي حيشتغل ان شاء الله.




## Finishing the Game

إنهاء اللعبة

When the kid finishes playing the game, please press the back button of your mobile. It looks like this: بعد ما الطفل يخلص لعب ياريت تدوس على زرار العودة بتاع الموبايل ودة شكله:



# Laila Shoukry, M.Sc.

CONTACT

**E-mail:** laila.shoukry@kom.tu-darmstadt.de

EDUCATION

PhD Researcher — December 2013 — 2019

Technical University of Darmstadt

Currently finishing my PhD thesis on the topic Mobile Multimodal Serious Games Analytics.

Master of Science in Digital Media Engineering and Technology (DMET) — 2012 — 2013

German University in Cairo, Egypt

- Grade: Excellent
- Supervised by Prof. Christian Sturm, Hamm-Lippstadt University, Germany and Prof. Galal H. Galal-Edeen, Amercian University in Cairo (AUC)
- Examined by Prof. Adel Khalifah and Prof. Fatma Moawad

Bachelor of Science with High Honors in Digital Media Engineering and Technology (DMET) — 2004 — 2009

German University in Cairo, Egypt

- Awarded full scholarship at GUC for the whole 5 years
- Ranked first on major in Bachelor
- Cumulative Grade: Excellent
- Bachelor Thesis wrote at Stuttgart University, Germany, as a result of a DAAD scholarship
- Supervised by Prof. Gunther Heideman and M.Sc. Sebastian Klenk at VIS, Stuttgart University
- Bachelor Thesis grade: Excellent
- After finishing Bachelor thesis, was awarded a scholarship for Masters at ULM University in Germany by DAAD

German Abitur Certificate — 1990 — 2004

DSB German School, Alexandria, Egypt

- Ranked 2nd on school
- Abitur Durchschnitt 1.2

WORK EXPERIENCE

AAST - Arab Academy for Science and Technology (Multimedia Department) — September 2017 — Present

Teaching Assistant

Part-time Teaching Assistant at the Multimedia and Computer Graphics Program at Computer Science Department.

KOM, TU Darmstadt, Germany — July 2015 — Present

Doctoral Student and Adjunct Research Assistant from abroad (Egypt)

Mini Fekra Learning Center — January 2018 — Present

Founder, Manager, Program Coordinator and Teacher

Creating mobile and online educational interactive mutlimedia and games as well as offering educational courses in different fields

Kinderland Education — December 2015 — December 2017

Program Coordinator and Teacher

KOM, TU Darmstadt, Germany — December 2013 — June 2015

Research Associate and Doctoral Student

Worked on several research projects at the university and gained experience in teaching and supervising student theses and projects in the research group Serious Games under the supervision of the research team leader Dr.Stefan Goebel.

CMA CGM Egypt for Navigation, Alexandria — August 2007 — September 2007

Internship at IT Department

German University in Cairo — 2006 — 2007

Junior Teaching Assistant

- Lab Assistance in Programming Lab for IET 4th semester in Spring 2006
- Teaching Assistance in Introduction to Computer Science for Pharmacy 2nd semester in Winter 2007
- Lab Assistance in Computer Graphics for CSEN 7th semester in Winter 2007

TRAINING

- June 2015 | Course on "Managing R&D Projects" at Ingenium at the Technical University of Darmstadt, Germany.
- Mai 2015 | Workshop on Presenting Research at Ingenium at the Technical University of Darmstadt, Germany.
- September 2014 | Course on "Basics of University Didactics: Organizing Tutorials and Lab Sessions" at the Department of University Didactics at the Technical University of Darmstadt, Germany.
- June 2014 | Seminar on "Time Management during the Phd process" at Ingenium at the Technical University of Darmstadt, Germany.
- June 2014 | Seminar on "Novel Time Management Techniques" at the Technical University of Darmstadt, Germany, organized by the In-House Women Traning Committee.
- Mai 2014 | Workshop on"Effective Communication in the Work Environment" at thTechnical University of Darmstadt, Germany, organized by the Women Promotion Working Group at the Mechanical Engineering Department
- Spring Semester 2013 | Teacher Training at Neue Deutsche Schule Alexandria (NDSA - German Future School Alexandria)
- July 2010 | First Aid Course at the Nursing College of Alexandria, Egypt.
- December 2005 | Workshop on "The Five Gates to Personal Achievement" organized by the GUC-SPSA at the German University in Cairo, Egypt.
- May 2006 | "Practical Training on how to present Islam" Program at Bridges Foundation, Cairo, Egypt.
- April 2006 | "Answering Questions and Refuting Misconceptions about Islam" Program at Bridges Foundation, Cairo, Egypt.
- March 2006 | "How to Present Islam" Program at Bridges Foundation, Cairo, Egypt.

AWARDS

- 2004 | Full Scholarship for the five years of study at the German University in Cairo
- 2006 | Certificate of Appreciation for being an active member in the academic community Pioneers AWG at the German University in Cairo

- 2007 | Cerificate of Appreciation for being the head of the Academic Committee at the Diggers Club for Computer and Information at the German University in Cairo.
- 2008 | Scholarship from DAAD for writing Bachelor Thesis at Stuttgart University
- 2008 | Scholarship from DAAD for pursuing M.Sc. at Ulm University (not done due to child care)
- 2012 | Report for AINAC Conference Competition about Game Hamza (Master Thesis Project) awarded Second Place in Entrepreneurship Category
- 2017 | Participation at the Qomra Competition with the Short Movie "The Box"
- 2018 | Participation at the ALECSO Competition in Tunisia with the Mobile Game "Hamza"

LANGUAGES

- **Excellent command of Arabic both spoken and written (Mother Language)**
- **Excellent command of English both spoken and written**.
  - [16.11.2016] IELTS Score: 8: Listening:9, Reading:8, Writing:8, Speaking:7, Overall: 8.
  - [22.06.2004] ELPET (The English Language Proficiency Entrance Test) at AUCairo, was eligible for ECLT 103/102.
  - [06.07.2004] TOEFL computer-based Exam Score 273 (=643 paper-based) – 28/30 in Listening, 30/30 in Structure/Writing, 24/30 in Reading
- **Excellent command of German both spoken and written.** (Abitur Grade 12 in German)
- **Very good command of French both spoken and written.** (Abitur Grade 14 in French)

## D Erklärungen laut Promotionsordnung

### §8 Abs. 1 lit. c PromO

Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

### §8 Abs. 1 lit. d PromO

Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

### §9 Abs. 1 PromO

Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

### §9 Abs. 2 PromO

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

*Darmstadt, 2020*

Laila Shoukry, M.Sc.