# Towards Better Tools to Support Mixed Methods in Game User Research

**Songjia Shen**
Games Studio
University of Technology,
Sydney
songjia.shen@uts.edu.au

**Chek Tien Tan**
Games Studio
University of Technology,
Sydney
chek@gamesstudio.org

**Tuck Wah Leong**
Interaction Design and Human
Practice Lab
University of Technology,
Sydney
tuckwah.leong@uts.edu.au

*CHI 2014 GUR Workshop*, April 26–May 1, 2014, Toronto, Canada.

## Abstract

There is an emerging body of research in mixing both qualitative and quantitative methods in game user research. However, differing nature of qualitative and quantitative methods makes analysis extremely hard, and there is insufficient research into designing better tools to support their effective combinations. This position paper provides a snapshot of current mixed methods in game user research and thereby presents several insights into designing software to support the effective combination of think-aloud and physiological data for a deep understanding of game user experiences.

## Author Keywords

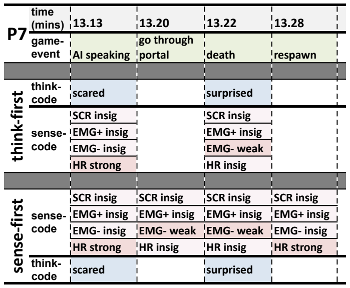
Game user research; psychophysiology; think-aloud

## ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces: Evaluation/methodology; I.2.1 [Applications and Expert Systems]: Games.

## Introduction

Both qualitative and quantitative methods play important roles in game user research (GUR). Although circumstantial differences sometimes dictate the appropriateness of one method over another,



| P7 | time (mins) | 13.13 | 13.20 | 13.22 | 13.28 |
|---|---|---|---|---|---|
| | game-event | AI speaking | go through portal | death | respawn |
| think-first | think-code | scared | | surprised | |
| | sense-code | SCR insig | | SCR insig | |
| | | EMG+ insig | | EMG+ insig | |
| | | EMG- insig | | EMG- weak | |
| | | HR strong | | HR insig | |
| sense-first | sense-code | SCR insig | SCR insig | SCR insig | SCR insig |
| | | EMG+ insig | EMG+ insig | EMG+ insig | EMG+ insig |
| | | EMG- insig | EMG- weak | EMG- weak | EMG- insig |
| | | HR strong | HR insig | HR insig | HR strong |
| | think-code | scared | | surprised | |

**Figure 1:** A small segment of P7's coded experience timeline. The gameplay is an example of a death sequence caused by P7 trying to navigate a puzzle involving a lethal energy ball that obstructs advancement. Refer to our main paper [4] for the full figure and analysis.

recent studies have shown that using combinations of various methods are beneficial in GUR (e.g., [1, 3, 5]). However, the fundamental differences in qualitative and quantitative data poses tremendous challenges. Qualitative methods like think-aloud and interviews provide rich data that are known to reflect reasonably accurate player states, but lack temporal precision, are often disruptive to the experience and require a lot of resources to gather and interpret. Quantitative methods like psychophysiology and telemetry provide high resolution data that are temporally precise and less disruptive, but lack contexts for interpretation and the sheer volume of data is daunting to process. Mixed methods hence possess a large potential to improving the understanding of player experiences.

Several researchers have explored the potential benefits of mixed methods in games, and at the same time demonstrated the difficulty in mixing disparate data. For example, Iacovides et al. [2] combined qualitative and quantitative data by collecting observations, post-play interviews, gaming diaries and physiological data. They found that the combination of multiple qualitative data sources provided useful insights as to where interesting game moments evolved as well as learning that occurred beyond gameplay. However, in the end they were unable to correlate the physiological data to the rest of the qualitative data as they had immense difficulties processing and visually interpreting the physiological signals. Another example is "biometric storyboards" [3], a tool that visualizes skin conductance responses (SCR) and facial electromyography (EMG) graphs with player self-report annotations. They showed that their tool can help yield a higher quality game than just using traditional playtests. However, their tool seemed to anchor on physiological data with the self-report comments added as auxiliary data, i.e., the tool leads the researcher in a certain analytic direction, which might incur some missing interpretations as we will highlight later in our work. Moreover, the comments are in their original form without reduction, and the researcher needs to mouse-over and consciously make sense of the text, which makes it hard to have a broader overview of the experience. The next section describes our efforts at unravelling some of the intricacies in combining think-aloud and physiological data, and hence try to address some of the issues above.

## Current Work

In our own work, we performed an in-depth study to investigate how to effectively combine think-aloud and physiological data. Full details can be obtained in our affiliated paper [4] and a brief overview is provided here.

We collected two main types of data, (1) video-cued retrospective think-aloud recordings and (2) physiological readings from nine participants playing the Portal game for 30 minutes. In order to obtain a deep understanding of their correlations and to investigate how to effectively mix them, we analysed the data twice, in different orders/directions. We coded both sets of data anchored on game events and performed one analysis 'led' by think-aloud data (think-first approach), and another analysis led by physiological data (sense-first approach).

In a nutshell, our findings show that many interpretations are lost when using only the think-first or only the sense-first approach. For example, we showed that solely using physiological signals as cues for locating think-aloud responses (sense-first approach) will result in many missed interesting experience reports from think-aloud. In general, we also reinforce the notion that using combined qualitative and quantitative data empowers the researcher/designer to better infer/interpret particular gameplay experiences. We also found that there were instances whereby think-aloud and physiological signals seemed to contradict, but can be reconciled by looking at larger temporal regions over a sequence of responses instead.

In conducting and analysing the data from our study, several key procedural issues also emerged. The main issue was the sheer amount of time required to convert the various sources of data into codes that could be placed on a visual timeline to be compared against (a snippet of an example experience timeline can be viewed in Figure 1). One issue was the task of synchronizing timestamps of the gameplay screen captures, the participant think-aloud videos and the physiological recordings.

For the physiological data, one time-consuming task was artefact rejection. For example, although we had already attached SCR sensors to the most stationary fingers whilst using a controller (middle phalanxes of ring and last finger), we still observed frequent significant artefacts due to finger movement. During the recording of the baselines for each participant, behavioral artefacts were also present, e.g., body fidgeting in anticipation to ‘get on’ with the game.

For the think-aloud data, it also took great efforts to reconcile the verbalizations to the gameplay video as the verbalization timings were often not well coordinated with the gameplay instances. For example, participants would often continue to talk about a particular instance in the video long after it has passed. In addition, the degree of verbosity and recall differed across participants. We sometimes had to regulate the think-aloud to make sure they are on track.

## Towards Better Tools

Based on the learnings from our study, we propose the following guidelines for building better tools to support the combination of think-aloud responses and physiological recordings.

1. *Interface should not lead the analysis direction.* This relates to our findings on how using one type of data to lead the other would result in lost information (i.e., think-first versus sense-first). For example, researchers should be careful not to over-rely on physiological signals to cue think-aloud responses, something that might appear to be intuitively useful. In general, it might be advantageous for researchers to perform multiple passes on multiple sources of data, especially when they are of a different nature (i.e., qualitative and quantitative). Hence the tool should be able to support multi-directional analyses, or at least not inhibit or lead the researcher in any specific analytic direction.

2. *Allow for many-to-many relationships between data types.* One prominent finding from our study was that multiple physiological codes can correlate to a single think-aloud code. This was related to our finding that think-aloud and physiological data can sometimes give us contradictory signals and inspecting multiple events temporally surrounding the anomaly often

helps to reconcile the interpretation. Hence the tool should facilitate mapping of multiple physiological codes to a single think-aloud code. The reverse (multiple think-aloud codes to a single physiological code) might also be possible although we did not observe it in our study.

3. *Automatically synchronize data to game events.* Though it appears straightforward, syncing the think-aloud recordings, the physiological recordings and the gameplay video should be an essential functionality of the tool. This allows for the interpretations to be anchored on the game events. An example of this functionality can already be seen in the biometric storyboards tool by Mirza-Babaei et al. [3].

4. *Automatically pre-process each data type.* From our study, the amount of time spent to pre-process each disparate data was a major limitation of using the mixed method approach for analysing player experiences. Hence the tool should aim to alleviate this chore as far as state-of-the-art computation techniques allow. For the physiological data, filtering and automatically converting the sensor data into codes should be straightforward as the data is purely quantitative. However, some thought should be put into automatically rejecting artefacts (e.g., using computer vision techniques to detect hand movements). For the think-aloud data, speech and lip recognition and technology can be used to aid the researcher in transcription and codification. However, this should only be used as a prompt to speed up the codification process, and not be overly reliant on its accuracy.

We hope that these guidelines have highlight the intricacies involved in using mixed methods in general, and that much research and thought needs to go into the design of tools to support such methods.

## Acknowledgments

This work was supported by the Centre for Human Centred Technology and Design and the School of Software at UTS.

## References

[1] Canossa, A., Drachen, A., and Rau Møller Sørensen, J. Arrrgghh!!!-Blending Quantitative and Qualitative Methods to Detect Player Frustration. In *Proc. FDG 2011*, ACM Press (2010).

[2] Iacovides, I., Aczel, J., Scanlon, E., and Woods, W. Making Sense of Game-Play : How Can We Examine Learning and Involvement ? *Transactions of the Digital Games Research Association 1*, 1 (2013), 1–17.

[3] Mirza-Babaei, P., and Nacke, L. How does it play better?: exploring user testing and biometric storyboards in games user research. In *Proc. CHI 2013*, no. May, ACM Press (2013), 1499–1508.

[4] Tan, C. T., Leong, T. W., and Shen, S. Combining Think-aloud and Physiological Data to Understand Video Game Experiences. In *Proc. CHI 2014*, ACM Press (2014).

[5] Zammitto, V., Seif El-Nasr, M., and Newton, P. Exploring Quantitative Methods for Evaluating Sports Games. In *Proc. CHI 2010 Workshop on Brain, Body and Bytes* (2010).