Towards a Serious Game Experience Model: Validation, Extension and Adaptation of the GEQ for Use in an Educational Context

De Grove Frederik
Ghent University
Korte Meer 7-9-11
9000 Ghent, Belgium
+32 9 264 84 76
frederik.degrove@ugent.be

Van Looy Jan
MICT-IBBT Ghent University
Korte Meer 7-9-11
9000 Ghent, Belgium
+32 9 264 84 76
j.vanlooy@ugent.be

Courtois Cédric
MICT-IBBT Ghent University
Korte Meer 7-9-11
9000 Ghent, Belgium
+32 9 264 84 76
cedric.courtois@ugent.be

ABSTRACT
In this paper, we present the results of game experience measurements of three design stages of the serious game Poverty Is Not a Game (PING) using the FUGA Game Experience Questionnaire (GEQ) extended with a Perceived Learning (PL) module. It is hypothesized that subsequent design stages will evoke a more positive game experience and higher PL. In a first step the factor structure and convergent and discriminant validity of the existing GEQ modules are tested yielding disappointing results. Next an adapted version is proposed yielding more acceptable results. Based on this model the different design stages are compared failing to yield significant differences either for most GEQ dimensions (except for challenge and competence which is probably related to usability issues) or for PL. Significant differences were found between classrooms however pointing to the importance of taking into account context in future research.

Categories and Subject Descriptors
K.8.0 [**Personal Computing**]: General - Games

General Terms
Measurement, Reliability, Human Factors, Standardization, Theory, Verification.

Keywords
Serious Games, Game Experience, Perceived Learning.

1. INTRODUCTION: SERIOUS GAME EXPERIENCE
Recent academic literature on gaming has seen a rise of interest in the idea of a measurable game experience [1] which is commonly conceptualized as that which makes gaming fun. A central concept in relation to enjoyable experiences is that of flow or optimal experience which can be described as "an optimal, intrinsically motivating experience induced by an activity in which one is fully absorbed" [2]. Apart from Sweetser & Wyeth [3] who attempted to fit this construct into a game-specific model with eight correlating dimensions (concentration, challenge, skills, control, clear goals, feedback, immersion and social interaction), attempts to thoroughly operationalise the multifaceted construct of game experience are scarce as most attempts are limited to using either one dimension of the game experience or employing a narrowed-down version of the flow concept [see e.g. 4, 5].

From 2006 to 2009, however the "Fun of Gaming" (FUGA) project funded by the European Community has worked towards measuring the human experience of media enjoyment. One of its core tasks was to develop a Game Experience Questionnaire (GEQ) to measure the game experience. This self-report measure consists of three modules: a core module, a social presence module and a post game module [6]. As far as we know, no attempt has been made as yet to validate the application of the GEQ to the domain of serious or educational games. In this paper we report on the testing of the alpha, beta and release candidate versions of the game PING.

Since the GEQ was primarily constructed to be used for measuring the experience of commercial games, its use in a formal learning environment requires the adoption of another experiential dimension, c.q. the experience of learning itself, which has in previous research been conceptualized as a flow effect [see e.g. 7]. The only attempt we know of to link perceived learning with game experience is that by Fu, Su and Yu [8] which consisted of a scale based on Sweetser & Wyeth's Gameflow Model [3] adapted to the specificity of what they call e-learning games. However, as this measure is only applicable to video games with clearly defined learning outcomes, it was decided that it was unfit to be used for a video game under development of which the primary aim was to raise consciousness about poverty.

It was therefore decided to build a perceived learning scale that takes into account the taxonomy used by Rovai et al. [9] (including affective and cognitive learning) and the first two levels proposed by Kirkpatrick [10] to evaluate training courses, namely Reaction (level 1) which pertains to the affect of the respondent towards the learning method and Learning (level 2) which corresponds with the affective and cognitive learning dimensions used by Rovai et al. [9].

As the GEQ was to be used outside the main context for which it was developed, our first research question was the following.

RQ1: Do empirical data confirm the construct validity of the GEQ?

Fun and Games 2010, September 15-17, 2010, Leuven, Belgium.

Moreover, it was expected that during the subsequent design stages, the game experience would evolve in a positive way.

H1: The game experience will become significantly more positive over the three design stages of PING.

Since a positive game experience and the experience of learning are intertwined, it is also expected that:

H2: There is a positive effect of the game experience on perceived learning.

And that:

H3: Perceived learning will rise significantly over the three design stages of PING.

2. THE GAME

PING (Poverty Is Not A Game) was commissioned by the King Baudouin Foundation and is an initiative as part of the European Year for Combating Poverty and Social Exclusion (2010). Its primary aim is to raise consciousness in adolescents concerning poverty and social exclusion in a way that is close to their everyday lives. The game takes place in a three-dimensional environment which represents an average Western European city. Players can choose between a male or female avatar. Although the decision to play with a certain avatar has an impact on the storyline, the central message the game wishes to convey stays the same. It hopes to raise consciousness concerning the mechanisms underlying poverty and is specifically aimed at what is sometimes referred to as the fourth world.

No detailed accounts exist concerning what has changed between the different design stages of PING (Alpha, Beta and Release Candidate). An informal conversation with the game developers revealed that changes primarily pertained to the story of the game and to the navigation in the game. The Alpha stage only consisted of a rudimentary storyline while orientation in the 3D game world was a challenge as no maps were readily available. In comparison, the Release Candidate (RC) provided a fully developed story which could be finished in about 50 minutes. Navigation was facilitated by a mini-map with GPS functionality. Similar in all design stages is the fact that PING had no sounds or music.

3. METHOD

3.1 Data Collection

Data were collected by testing PING in a total of 22 classrooms. To measure game experience, an online survey consisting of several blocks was used. A first block was the core module of the GEQ and inquired how the player felt whilst playing the game (92 five-point Likert scale (FPLS) items: Not at all, Slightly, Moderately, Fairly and Extremely). A second block was the social presence module and consisted of 25 FPLS items. Then a first part of the perceived learning scale followed (10 FPLS items). Next are the post-game module (21 FPLS scale items) and the second part of the perceived learning scale (10 FPLS items). A subsequent block of the survey explores the gaming behaviour of the respondent. The three final questions asked about socio-demographic characteristics (gender, age and education level). In total, the survey was filled out by 373 respondents of which 340 were retained after data cleaning.

3.2 Scale Validation

We first assessed the construct validity of the GEQ to evaluate its appropriateness in the context of a pre-launch serious game in a formal learning environment.

Construct validity was accounted for by testing the factor structure of the core module, the social presence module and the post game module and by using measures accounting for convergent and discriminant validity. To test the factor structure we made use of confirmatory factor analysis which "*is a way of testing how well measured variables represent a smaller number of constructs*"[11]. As proposed by Hair et al. [11] we checked convergent validity by means of the coefficient alpha and the average percentage of variance extracted (VE). Discriminant validity was examined by comparing the VE of two constructs with the squared correlation of those two constructs. The rationale behind this is that a latent construct should explain the variance in its items better than that it explains another construct. To complement the GEQ a perceived learning scale was constructed and tested for its construct validity.

4. RESULTS

4.1 Construction of a Serious Game Experience Model

4.1.1 Core Module

Goodness-of-fit indices (N=330, $\chi 2/df = 2.86$, CFI = .78, TLI = .75, RMSEA = .075, CI90 = .071, .080) were not satisfactory as CLI and TLI scores suggest that the proposed model did not fit our data. This is confirmed when checking for convergent and discriminant validity as none of the core module's dimensions has an acceptable VE. This "*indicates that on average, more error remains in the items than variance explained by the latent factor structure imposed on the measure*" [11] which points to a problematic scale. When checking for discriminant validity, six concepts show considerable similarity with other concepts, i.e. Competence, Immersion, Flow, Positive Affect, Challenge and Annoyance. More specifically Competence, Immersion and Flow explain each other better than they explain the variation in their own items. The same is true for Annoyance and Challenge and for Positive Affect and Competence. Only the concept Negative Affect proves to be different enough.

4.1.2 GEQ Social Presence Module

Goodness-of-fit indices of the social presence module were similar to that of the core module (N=330, $\chi^2/df = 3.68$, CFI = .84, TLI = .81, RMSEA = .090, CI_{90} = .081, .099), which indicates that the data do not fit the proposed model. Moreover none of the dimensions has an acceptable VE statistic while coefficient alphas seem to suggest reliable scales. Discriminant validity was not satisfactory for Empathy and Behavioural Involvement and for Negative Affect and Behavioural Involvement.

4.1.3 GEQ Post Game Module

Goodness-of-fit statistics show that our data do not fit the proposed model (N=330, $\chi 2/df = 4.44$, CFI = .84, TLI = .81, RMSEA = .102, CI90 = .093, .111). Concerning convergent validity, all dimensions yielded an acceptable coefficient α but only Tiredness explained sufficient variance. Discriminant analysis for the post game module showed that three of the four

dimensions are not different enough. It concerns Negative Experience which relates to Tiredness and Returning to Reality.

4.1.4 Construction of a Perceived Learning Scale

Of the total item pool (20 items), nine items were retained to construct the perceived learning scale. Two items composed the construct of Affective Gaming which assesses how one responds to receiving education through video games while the construct of Learning (7 items) explores to what extent the respondent thinks they have learned something on the topic of poverty. This model was inputted in AMOS and resulted in a good fit (N=330, $\chi 2/df$ = 1.79, CFI = .99, TLI = .98, RMSEA = .049, CI90 = .025, .071) while both Convergent and Discriminant validity proved to be acceptable.

4.1.5 Towards a Serious Game Experience Model

As the current structure of the GEQ was considered as inadequate for our further exploration (**RQ1**), it was decided to build our own serious game experience model in which the item pool and dimensions of the GEQ served as a starting point. Furthermore, we aimed to incorporate our own perceived learning scale. The first decision we made was to leave out the social presence module since this module was not deemed fit to assess the complex and rich nature of social interactions that emerge during gameplay in a formal learning environment. Moreover, the construct of Tiredness of the post game module was omitted as it proved to be highly context-dependent (being tired was not due to gaming but to the moment of testing).

All three modules were reviewed and adapted based on item distributions, coefficient alphas, corrected item-total correlations and the one-dimensionality of the dimensions. This eventually resulted in a model in which game experience and post game experience were conceptualized as second order constructs. Game experience is composed of eight first order constructs: Competence (2 items), Vividness (3 items), Negative Affect (2 items), Positive Affect (3 items), Immersion (2 items), Challenge (2 items), Affective Gaming (2 items) and Learning (7 items). Post game experience consists of Positive Experience (3 items), Negative Experience (3 items) and Returning to Reality (2 items). Annoyance was omitted since this construct proved to be impossible to build from the applicable data. Furthermore, the construct of Challenge did not result in an acceptable scale either but was retained on the basis of theoretical and practical considerations. Fit indices of the proposed Serious Game Experience model yielded an acceptable fit (N=330, $\chi 2/df$ = 1.84, CFI = .93, TLI = .92, RMSEA = .050, CI90 = .045, .056). This model served as the foundation for our further exploration of the serious game experience of PING during its subsequent design stages. At this moment, it is interesting to note that Learning has a standardized regression weight of .50 which confirms our hypothesis that there is a positive effect of the game experience on perceived learning (**H2**).

4.2 The Evolution of Game Experience

4.2.1 During Subsequent Design Stages

To compare how the different dimensions behaved during subsequent design stages, we analyzed the variance within groups and between groups (ANOVA, power = .98, Effect size = .025).

Results indicated that, over the three different design stages, only Competence ($p < .005$, $F = 6.03$, $df = 335$) and Challenge ($p < .001$, $F = 5.37$, $df = 330$) differed significantly. Post-hoc tests (Scheffe) show that these differences are to be found between the Alpha stage and the RC stage. This applies to Competence ($p < .003$) as well as to Challenge ($p < .021$). On average, Competence scores were lower during Alpha testing (M = 2.73, S.D. = .13) than during RC (M = 3.26, S.D. = .09) testing while scores for Challenge were higher for Alpha testing (M = 2.05, S.D. = .11) compared to RC testing (M = 1.72, S.D. = .07).

Learning is marginally significant ($p < .065$, $F = 2.80$, $df = 329$) but differences for the design stage are situated between Alpha and Beta where average Beta scores (M = 2.95, S.D. = .08) were higher than Alpha scores (M = 2.79, S.D. = .09).

4.2.2 Classroom Comparison

To check if our sample size was big enough to execute an ANOVA with 22 groups, we performed a power analysis. With a power of .77 our data will only be capable to reliably detect large or, to a lesser degree, medium differences. Notwithstanding our relatively low power, a considerable number of dimensions proved to differ significantly between classrooms (Competence, $p < .001$, $F = 2.317$, $df = 335$; Vividness, $p < .012$, $F = 1.880$, $df = 330$; Challenge, $p < .020$, $F = 1.784$, $df = 330$; Negative Affect, $p < .000$, $F = 3.428$, $df = 338$; Positive Affect, $p < .000$, $F = 2.684$, $df = 335$; Affective Gaming, $p < .000$, $F = 2.234$, $df = 329$; Learning, $p < .011$, $F = 1.900$, $df = 329$; Negative Experience, $p < .000$, $F = 2.642$, $df = 329$). Only Immersion, Positive Experience and Returning to Reality do not differ. Regrettably, we did not have enough data at our disposal to identify differences between individual classrooms. Although we only had 22 classrooms (level 2 units) it was decided to perform a multilevel analysis to further explore if unexplained variance could be found on classroom level. We found significant differences in the variation of intercepts for Positive Affect ($p < .05$, $F = 1176$, $df = 19$) and Negative Affect ($p < .05$, $F = 431$, $df = 17$). Calculating the intra-class correlation coefficient resulted in 10% of unexplained variance on level 2 for Positive Affect while Negative Affect had 15% of unexplained level 2 variance. Considering our small number of level 2 units, these results can be considered as a further indication of classroom effects.

5. CONCLUSION / DISCUSSION

The testing of PING yielded some remarkable results. First, there is the fact that only Competence and Challenge differ significantly between the Alpha and RC stage. When we add the fact that, first of all, variation in Challenge was not explained by game experience ($R^2 = .01$, $p < .861$) but shared unexplained variation with Competence ($r = -.57$), and secondly, that one of the two major changes during the design stages pertains to the usability of the game (navigation), some interesting assumptions can be made. On theoretical grounds, both Competence and Challenge can be connected to usability issues. Usability can be considered as a prerequisite for a good game experience but it is not equal to it. As such, it is possible that Challenge is actually a measure of usability which would explain why it did not fit our Serious Game Experience Model. Furthermore, this could also explain some of the error variance of Competence. As such, we did not find a significant positive change in the game experience during the different design stages (**H1**), but we did find a significant change in the experienced usability of PING.

Although marginally significant, perceived learning changed between the Alpha and Beta stages of the game. This is probably

due, however, to the fact that the Beta stage had some atypical distributions. When checking for interaction effects with Gender of Educational level by means of a multivariate analysis the difference in Learning ceased to be (marginally) significant. The fact that perceived learning does not change positively during the subsequent design stages (**H3**) is surprising. Especially because the storyline was one of the major changes in the design flow (cf. supra). Furthermore, students were allowed to play the Beta and RC versions longer than the Alpha version which could have resulted in a better learning experience. On the other hand, this finding is not illogical if we take our serious game experience model into account. As most of the dimensions do not vary between the different design stages, it is logical that learning does not vary either. More specifically, if the storyline would have changed enough, this would have been reflected in the concept of Vividness. Consequentially, the experience of perceived learning would have changed too. This indicates that the changes in the subsequent design stages of PING were not large enough to evoke an improved learning experience. Considering the pre-launch status of the game, an interesting starting point could have been to use the GEQ in combination with a validated usability measure. That way, improvements in usability could have been linked to game experience. As such, it would be interesting if future research on video games that are under development would incorporate usability as well as game experience measures.

Finally, it is interesting to see that there seem to exist strong differences between classrooms on most of the game experience dimensions. Perhaps the most remarkable result is that, when gaming, only Immersion does not differ significantly between classrooms. A possible explanation could be that social interaction during gameplay prevents Immersion to go above a certain level while the absence of sounds or music could have been a decisive factor in stimulating social interaction. Equally intriguing is the fact that constructs such as Competence, Vividness, Challenge and Learning seem to have a collective component. With our current dataset, however, we could not explore this further. Future research could consider using focus groups to explain these findings. Another approach might be to use an experimental design in which the content-related variable is manipulated.

As such, using the GEQ for a serious game in a pre-launch status in a classroom context might not be ideal as some of the items are highly context dependent (Flanders versus the Netherlands, education level of respondents, etc.). However, our own proposed model is susceptible to criticism as well. With the exception of Learning, the constructs of our model contain only two or three items. This is barely enough to cover the concepts they intend to measure hence resulting in operational narrowing of the concepts involved. Results also indicate that the concept of Challenge does not fit its theoretical content (conceptual displacement). Moreover, we built the model from the available data. Since our data ensued from one video game only, the possibility exists that our model cannot be generalized. Further testing will reduce uncertainty. However, future attempts to construct a game experience model should try to include concepts such as Concentration, Feedback, Clear Goals and Control while a model aimed at measuring experiences in a classroom context should try to incorporate a measure that is able to take context effects into account such as the rich diversity of social interactions.

6. REFERENCES

[1] IJsselsteijn, W., de Kort, Y., Poels, K., Jurgelionis, A. and Bellotti, F. Characterising and measuring user experiences in digital games. ACM, City, 2007.

[2] Csikszentmihalyi, M. Flow: The psychology of optimal experience. Harper & Row New York, 1990.

[3] Sweetser, P. and Wyeth, P. GameFlow: a model for evaluating player enjoyment in games. Computers in Entertainment (CIE), 3, 3 2005), 3.

[4] Douglas, J. and Hargadon, A. The pleasures of immersion and engagement: schemas, scripts and the fifth business. Digital Creativity, 12, 3 2001), 153-166.

[5] McMahan, A. Immersion, engagement and presence. The video game theory reader2003), 67-86.

[6] Poels, K., De Kort, Y. and IJsselsteijn, W. Measuring the human experience of media enjoyment. Technical University, City, n.d.

[7] Kiili, K. Digital game-based learning: Towards an experiential gaming model. The Internet and higher education, 8, 1 2005), 13-24.

[8] Fu, F., Su, R. and Yu, S. EGameFlow: A scale to measure learners' enjoyment of e-learning games. Computers & Education, 52, 1 2009), 101-112.

[9] Rovai, A., Wighting, M., Baker, J. and Grooms, L. Development of an instrument to measure perceived cognitive, affective, and psychomotor learning in traditional and virtual classroom higher education settings. The Internet and higher education, 12, 1 2009), 7-13.

[10] Kirkpatrick, D. Evaluating training programs: The four levels. Berrett-Koehler, 1998.

[11] Hair, J., Black, W., Babin, B., Anderson, R. and Tatham, R. Multivariate Data Analysis. Upper Saddle River, NJ: Prentice Hall, City, 2006.