Journal of Educational Multimedia and Hypermedia (2015) **24**(1), 43-74

Development and Validation of the Game Perception Scale (GPS)

SYLKE VANDERCRUYSSE, MIEKE VANDEWAETERE,
AND MARIE MAERTENS
KULeuven Kulak, Belgium
sylke.vandercruysse@kuleuven-kulak.be
Mieke.Vandewaetere@kuleuven-kulak.be
Marie.Maertens@kuleuven-kulak.be

JUDITH TER VRUGTE
University Twente, Netherlands
j.tervrugte@utwente.nl

PIETER WOUTERS
University Utrecht, Netherlands
P.J.M.Wouters@uu.nl

TON DE JONG
University Twente, Netherlands
A.J.M.deJong@utwente.nl

HERRE VAN OOSTENDORP
University Utrecht, Netherlands
H.vanoostendorp@uu.nl

JAN ELEN
KuLeuven, Belgium
jan.elen@ppw.kuleuven.be

Despite the pervasiveness of perception and considerable impact of perception on the use of ICT for educational purposes, there is a surprising paucity of perception assessment instruments. The present proposal expands on this through

the development and initial validation of the Game Perception Scale (GPS). Based on perception literature, perception is defined as (1) students' expectations about the goals of the environment and (2) the degree to which a student believes that using game based learning environments will enhance his or her performance on which the game based learning environment focuses. In a first study, the exploratory factor analyses revealed a meaningful two-factor solution, which reflects the two dimensions that constitute the notion of game perceptions. Further, we used a confirmatory factor analysis (CFA) to validate and confirm the model. Cross-validation was repeated in second part of this article in which a large sample was used to perform multi group CFAs. The results showed that GPS can be used in different target groups when researchers want to measure students' perceptions of game-based learning environments for both pro- and retrospective purposes. However, significant differences in structural relations with respect to the covariance and the variance of the perceived use and perceived goal subscale were found for all the groups.

Interactive technologies, like educational games, are increasingly expected to be effective and efficient tools for supporting the acquisition of complex subject matter (Ke, 2008; Ke, 2009; Ritzhaupt, Higgins, & Allred, 2011). They are expected to evoke intense engagement and motivation in the learning process (e.g., O'Neil, Wainess, & Baker, 2005; Vogel, Greenwood-Ericksen, Cannon-Bowers, & Bowers, 2006), to actively involve students in challenging situated problem solving (e.g., Becker, 2007; Garris, Ahlers, & Driskell, 2002), to enhance learning and understanding (Hayes & Games, 2008) and to improve student's performance (Liu & Chu, 2010). Notwithstanding the popularity of educational games in education, empirical research and evidence for the claims and expectations remain limited (Kebritchi, 2008; Papastergiou, 2009; Wouters, van der Spek, & van Oostendorp, 2009) resulting in a gap between what is theoretically claimed and what has been empirically demonstrated. This is a.o. due to the lack of a univocal definition or a shared framework that can be used to investigate the effects of educational games (Vandercruysse, Vandewaetere & Clarebout, 2012). Additionally, research on the effect of games is complicated because game based learning (GBL) is not a stand-alone or isolated activity (Hays, 2005). Multiple aspects need to be considered such as the environment (e.g., type of game, domain, etc.), the learning results and mediating variables (e.g., learner characteristics, etc.). Investigations on mediating

variables reveal that direct effects of instructional interventions (e.g., playing an educational game) are not likely to be expected (Lowyck, Elen, & Clarebout, 2004). One type of mediating variables, as proposed by the cognitive meditational paradigm of Winne (1987), is students' cognitions. These cognitions (such as perceptions and beliefs) have a profound impact on how the learner interacts with the learning environment, which is then related to learning outcomes. Hence, knowing students' cognitions might help to gain insight in the widespread effects of educational games. In the following part we will focus on one specific aspect of students' cognitions, namely perceptions.

Theoretical Background – Why Taking Perceptions into Account?

The importance of students' perceptions was already demonstrated by a studie of Salomon (1984). Salomon (1984) demonstrated that students' differential learning may depend on what they perceive the material (i.e. video lessons) to be. If television was perceived as being part of 'easy' leisure time, a way to relax, invested mental effort was lower compared with perceptions of television as an instructional medium. Taking students' perceptions about the materials' affordances into account was of added value in understanding the relationship between the materials and students' learning processes (i.e., the investment of processing effort). This was also corroborated by Winne (1987). In the cognitive meditational paradigm, Winne (1987) stresses the importance of learners' perceptions when studying the relationship between interventions and learning. Although an instructional intervention can be designed to be very powerful (e.g., contain a well-defined learning goal), students' perceptions of this intervention and subsequent learning goal will determine what kind of learning activities will be employed. Correspondence between students' perception of the intervention - their expectations about the goals of the intervention - and the intention of the intervention can optimize the effects of the intervention (Vandewaetere, Vandercruysse, Clarebout, 2012). An example is the use of a game based learning environment (GBLE). If the goal of the intervention is to enhance students' math performance by using a GBLE, but students perceive the GBLE to be a leisure activity, there is a discrepancy between designers' or teachers' intentions, and students' learning goals, which is likely to result in a less effective learning process and, hence, in a less effective GBLE.

Another framework in which students' perception is considered is the Technology Acceptance Model (TAM). TAM focuses, amongst other things (e.g., perceived ease of use), on the perceptions of the usefulness of tech-

nology (Davis, 1989), and is defined as 'the degree to which a person believes that using a particular technology would enhance his or her (job) performance' (Davis, 1989, p. 320). Perceived usefulness has been shown to influence the acceptance and the intention of students to use the technology (i.e. how and when they will use the technology). When extending this definition to the use of a GBLE, perceived usefulness can be described as the degree to which a person believes that using a GBLE would enhance his or her performance. This will influence in turn how and when they will use the GBLE.

Taken together, knowledge of how students perceive a given instructional intervention seems essential because inter-individual differences in perceptions may affect learning results and the effectiveness of the intervention (Lowyck, et al., 2004; Struyven, Dochy, Janssens, & Gielen, 2008). How students perceive instructional interventions triggers their engagement in learning, and mediates the effect of the instructional interventions on their learning (Elen & Lowyck, 2000; Entwistle, 1991; Lowyck et al., 2004; Shuell & Farber, 2001). Therefore, students' perception may have a more significant impact on students' outcomes than the learning environment in itself (Shih, & Chuang, 2013). In this study the focus is on students' perceptions because the effectiveness of GBL may be largely affected by students' perceptions about the GBLE. Therefore it is important that researchers are able to reliably measure students' perceptions. The development and validation of the Game Perception Scale (GPS), as presented in this research, serves this goal.

Defining Perceptions for GBL - Scale Construct and Content

Considering the importance of students' perceptions, and the gap in previous GBL literature to measure game perceptions as defined above, we want to develop and validate the Game Perceptions Scale (GPS; see Table 1). From previous mentioned literature, we deducted two key aspects of game perception, namely: (1) students' expectations about the goals of the GBLE (perceived goal) and (2) the degree to which a student believes that using a GBLE will enhance his or her performance (perceived usefulness). Both aspects are represented in a subscale of the GPS.

Perceived Goal (PG). In line with Salomon's (1984) findings, we want to know whether students perceive a GBLE as a learning or a playing experience, the first set of questions assesses the perceptions of the individual players about the goal of the GBLE. The items more concretely focus on the distinction between learning and playing and are based on the 'direct play

assessment' subscale of the play experience scale (PES; Pavlas, Jentsch, Salas, Fiore, & Sims, 2012). With the PES, Pavlas et al. (2012) assess the subjective experience of play. The subscale 'direct play assessment' focuses on the participants' engagement in play by directly referring to the opposition between work and play. Because our focus is on learning instead of working, this was adjusted. The items as used for the PG subscale are shown in Table 1.

Perceived Usefulness (PU). The second set of items in the GPS assesses the perceived usefulness of using a GBLE. These items are based on the usefulness subscale of the Intrinsic Motivation Inventory (IMI; McAuley, Duncan, & Tammen, 1987), since this scale includes a motivational component which is not made explicit in the PU subscale of the TAM. This latter subscale is more focusing on the perceived costs and benefits, and thus on effectiveness, productivity, time savings, etc. (e.g., using a GBLE enables me to accomplish tasks more quickly, using a GBLE increases my productivity, etc.). Since our interest is more on the relation between students' perceptions and their motivation to interact with a GBLE, the value/usefulness subscale of the IMI is selected. This subscale addresses more generally the idea of using a GBLE during the learning processes, focusing on students' perceptions about the usefulness of using a GBLE. This subscale was used in many studies (in combination with other IMI subscales), like for instance in a study of internalization with an uninteresting computer task (Deci, Eghrari, Patrick, & Leone, 1994). Only items that explicitly focus on the learning or playing aspect were included. For example, items 6 and 7 (as presented in Table 1) were not included in the PU subscale that is used for this study because they are to generally formulated and not specific enough for our context (i.e. making use of a GBLE).

Table 1

Original Subscales Items (Source of Inspiration for GPS) and GPS Subscales Items

N°	Original subscale (original Questionnaire)	GPS Subscale	GPS Item
1	Value/ usefulness (IMI)	PU	I believe this activity could be of some value to me.
2	Value/ usefulness (IMI)	PU	I think that doing this activity is useful for _____
3	Value/ usefulness (IMI)	PU	I think this is important to do because it can help me___
4	Value/ usefulness (IMI)	PU	I would be willing to do this again because it has some value to me.

Table 1 continued

N°	Original subscale (original Questionnaire)	GPS Subscale	GPS Item
5	Value/ usefulness (IMI)	PU	I think doing this activity could help me to _____
6	~~Value/ usefulness (IMI)~~	PU	~~I believe doing this activity could be beneficial to me.~~
7	~~Value/ usefulness (IMI)~~	PU	~~I think this is an important activity.~~
1	Direct play assessment (PES)	PG	When I was using the game, it felt like I was playing rather than working.
2	Direct play assessment (PES)	PG	I would characterize my experience with the game as 'playing'.
3	Direct play assessment (PES)	PG	I was playing a game rather than working.
4	Direct play assessment (PES)	PG	Using the game felt like work.

Literature indicates that many variables, for example perceptions, influence technology use (Davis, 1989) and thus are present even before they get the chance to use the technology, in this case GBLE. This perception is created based on prior experiences of students with games or GBLEs. Therefore, students' perception is measured both before and after game play. The pre-measurement will give an indication about how students think about GBL before they even got the chance to play the game. In other words, students' expectations about the usefulness and goal of playing the educational games are measured. The post-measurement provides information about students' perceptions after the gameplay, and allows researchers to investigate whether the interaction with the GBLE was of any effect on students' perceptions. In Table 2 the items as used in the validation studies are presented. Responses to the items of the GPS were measured by a 6-point Likert scale ranging from strongly disagree (1) to strongly agree (6), in line with the PES questionnaire.

Table 2

GPS Subscales Items as presented to the students in our studies

Item n°	Subscale	Item
Item 1	PU	I believe playing this game could be of some value to me.
Item 2(R)	PG	Using the game will feel like learning proportional reasoning.
Item 3	PU	I think playing this game is important to do because it can help me to learn proportional reasoning.
Item 4	PG	I will be playing a game rather than learning proportional reasoning.
Item 5	PU	I think playing this game could help me to learn proportional reasoning.
Item 6	PG	I will characterize my experience with the game as 'playing' rather than learning.
Item 7	PG	When I will use the game, it will feel like playing rather than learning.
Item 8	PU	I would be willing to play this game because it has some value to me to learn proportional reasoning.
Item 9	PU	I think that playing this game is useful for learning proportional reasoning.

Note. R= item is formulated in reversed order and should be reversed before data can be analyzed.

These items are based on a GBLE in which students learn to solve proportional reasoning problems.

Scale Validation: Aim of the Studies

Having defined the construct of game perceptions in the introduction, the following part focuses on the validation of the GPS. The validation procedure followed different steps, as suggested by Vandewaetere and Desmet (2009). First an exploratory factor analyses (EFA) was conducted. Using this analysis we tried to identify and corroborate the two underlying factors (i.e., PU and PG) of game perceptions. In the next step, a confirmatory factor analysis (CFA) was used for validating and confirming the factor structure as obtained by the EFA (Matsunaga, 2010). To investigate the suitability of the GPS in different settings, (i.e. different target groups, different GBLEs and different amount of playtime) a multiple group CFA was done. Data from three different studies was used to generalize the valida-

tion and to investigate to what extent the subscales of the GPS are invariant (i.e. equivalent) across these particular groups (Byrne, 2001). A first distinction we made was between pupils from primary vs. secondary school students. Previous research showed that measuring learner characteristics like perceptions, with children, is not always that reliable. Having an insight for which age groups the instrument is valid, seems desirable. Secondly, we distinguished between two different GBLEs. Because of the wide range of GBLEs available and the rapidly changing landscape of games (de Freitas, 2006) having an instrument that remains valid for different GBLEs is convenient. The last distinction concerns the duration of gameplay. Implementing games in classrooms remains teachers responsibility although it remains unclear for teachers how much and in which ways games need to be used most effectively to support learning (Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; de Freitas, 2006). Because of this indistinctiveness, and also due to the lack of time available to implement these games (Kirriemuir & Mc Farlane, 2004) the duration of the implementation of games in class often differs. Having an instrument that is not sensitive for this variation is advisable.

PART 1: EXPLORATORY AND CONFIRMATORY FACTOR ANALYSIS

In this study, the psychometric quality of the GPS was tested via a study in which students got the chance to play with a GBLE.

Method

Participants

298 vocational track students participated in this study. Participants were students from 3th and 4th year of secondary education, with an age range between 14 and 17 years. All students were recruited from 14 classes in 2 secondary schools.

GBLE

The GBLE named ‘Zeldenrust’ is a self-developed educational game that students play individually (Vandercruysse, ter Vrugte, Wouters, Clarebout, Elen, 2012). It concerns a 2D cartoon-like environment meant for

14-16 year old vocational students. The game focuses on mathematical problems, more specifically proportional reasoning problems. This content is part of the students' curriculum. Since teachers in vocational education mention that their students often experience difficulties with proportional reasoning, this content is also practically relevant. Because the GPS was also filled in before students got the chance to play the game, expectations about the GBLE were measured rather than experiences with the environment.

Procedure

Before and after gameplay the GPS questionnaire was completed by the participants. The study was organised as part of students' mathematical class. Students were informed by teacher that they got the chance to play this math game during two hours instead of having regular mathematics lessons. During gameplay, the teacher and researcher were at students' disposal.

Results

Exploratory factor analysis: principal component analysis

Descriptives. For this analysis, only pre-measurement of GPS was taken into account. The GPS was completely filled in by 248 students. In Table 3 the means and standard deviations for the items are shown. As shown in the table below, most mean scores range between 3 and 4 except for item 4 ('I will be playing a game rather than learning proportional reasoning') which has a score larger than 4. A mean score on each item higher than 3, indicates that the participants have rather positive perceptions about the goals and usefulness of the game to play, i.e. by playing the game, students belief that learning proportional reasoning can be enhanced. Based on these answers a principal component analysis (PCA) was conducted and the factor hypotheses was tested with a CFA.

Table 3
Descriptive Statistics for the 9-item GPS

Number	Item	*n*	*M*	*SD*
Pre Item 1	I believe playing this game could be of some value to me.	248	3,68	1,32
Pre Item 2 R	Using the game will feel like playing and not like learning proportional reasoning.	248	3,41	1,35
Pre Item 3	I think playing this game is important to do because it can help me to learn proportional reasoning.	248	3,61	1,26
Pre Item 4	I will be playing a game rather than learning proportional reasoning.	248	4,38	1,50
Pre Item 5	I think playing this game could help me to learn proportional reasoning.	248	3,63	1,25
Pre Item 6	I will characterize my experience with the game as 'playing' rather than learning.	248	3,34	1,31
Pre Item 7	When I will use the game, it will feel like playing rather than learning.	248	3,48	1,26
Pre Item 8	I would be willing to play this game because it has some value to me to learn proportional reasoning.	248	3,24	1,37
Pre Item 9	I think that playing this game is useful for learning proportional reasoning.	248	3,63	1,32

Reliabiltiy

Reliability was tested by calculating the internal consistency using Cronbach's alpha (Cronbach, 1951). The value of alpha depends on the number of items and the dimensionality of the questionnaire and on the intercorrelation of the items (Cortina, 1993). Taking the number of items into account (i.e. $n = 9$), the supposed dimensionality (i.e. 2 factors) as well as average item intercorrelations (i.e. $r = .28$), the alpha value for the pre-GPS ($\alpha = .77$) can be considered as very good (Cortina, 1993). The reliability for PU subscale was $\alpha_{pu\text{-}pre} = .88$ which is good according to the number of items (i.e. $n = 5$), the unidimensionality (i.e. 1 factor) and the average item intercorrelation (i.e. $r_{pre} = .59$) (Cortina, 1993). The PG subscale, however, showed a lower reliability level ($\alpha_{pg\text{-}pre} = .67$), taking into account the number of items ($n = 4$), the unidimensionality (i.e. 1 factor) and the average item intercorrelation (i.e. $r_{pre} = .35$) (Cortina, 1993). Table 4 gives an overview per item of the item total correlations and of Cronbach's alpha if the item was deleted. Item 2 and 4, which are items of the PG subscale seems

to be problematic items since they have a low item total correlation and the value of Cronbach's alpha increase after deleting the items. This might be an explanation for the low reliability value of this subscale.

Table 4

Item total correlation per item and Cronbach's alpha if item was deleted

	Item total correlation	Cronbach's alpha if item deleted
Pre Item 1	,491	,738
Pre Item 2 R	,166	,786
Pre Item 3	,608	,720
Pre Item 4	,176	,789
Pre Item 5	,597	,722
Pre Item 6	,460	,742
Pre Item 7	,450	,744
Pre Item 8	,578	,723
Pre Item 9	,588	,722

Principal component analysis (PCA)

A PCA was conducted on the 9 items with oblique rotation (Promax) because it is possible that the underlying factors (PG and PU) correlate with each other (Brown, 2009; Field, 2009). The Kaiser-Meyer-Olkin (KMO) measure verified the sampling adequacy for the analysis (Field, 2009). The *KMO* = .82 is very good (Hutcheson & Sofroniou, 1999). The *KMO* values for individual items were > .68 which is good. Barlett's test of sphericity (λ^2_{pre} (36) = 842.64, $p < .001$) indicated that correlations between items were sufficiently large for PCA (Field, 2009). On the basis of these statistics, we deemed factor analyses appropriate.

An initial analysis was run to obtain eigenvalues for each factor in the data. The GPS data indicated two factors with eigenvalues over Kaiser's criterion of 1(Kline, 2011) which in combination explained 61.10% of the variance. The scree plot showed inflexions (Field, 2009) that would justify retaining two factors. Given the results of the analysis and the scree plot of the GPS, two factors are retained. Table 5 shows the factor loadings after rotation, the eigenvalues and the amount of explained variance.

The items that cluster on the same factor (>.70) suggest that factor 1 in both questionnaires represents the PU-dimension and factor 2 more or less represents the PG-dimension. Items 2 and 4 seem to be exceptions, for which item 2 shows insufficient load with the PG-dimension as expected. Item 2 ('Using the game will feel like playing and not like learning proportional reasoning') contributed least to the factor solution. It was the only item that was formulated in reversed order and therefore might be confusing for our target group. Additionally, Table 4 showed a low item total correlation of item 2 and an improvement of the Cronbach's alpha after item 2 was deleted (α = .79). For the PG-subscale, the value of Cronbach's alpha improved ($\alpha_{PG\ pre}$ = .70) after deleting item 2, but did not improved after deleting item 4 ($\alpha_{PG\ pre}$ = .67). Therefore only item 2 was deleted from the questionnaire.

Table 5

Factor loading after rotation, eigenvalues and amount of explained variance. Numbering of items is in accordance with the numbering in Table 2

	Component	
	1	2
GPS Pre Item 3	.856	-.006
GPS Pre Item 9	.854	-.029
GPS Pre Item 5	.844	-.014
GPS Pre Item 8	.824	.007
GPS Pre Item 1	.728	-.023
GPS Pre Item 6	.086	.827
GPS Pre Item 7	.081	.817
GPS Pre Item 4	-.126	.648
GPS Pre Item 2R	-.096	.571
Eigenvalue	3.555	1.944
% of explained variance	39.500	21.598

Confirmatory factor analysis (see Footnote 1)

Data preparation

For this analysis, again only pre-measurement of GPS was taken into account. The complete sample of participants was randomly divided in two

samples: a calibration sample (n = 149) and validation sample (n = 149). The calibration sample data were used to validate and confirm the factor structure (as obtained by the EFA) by means of a CFA and the validation sample was used to cross-validate the solution obtained with the calibration sample (Vandewaetere & Desmet, 2009).

Missing data was deleted listwise (n =50) and outliers (n = 6) were detected using the Mahalanobis D^2-value which was compared with the critical value (i.e. 26.13; Kline, 2011). This resulted in 119 participants in the calibration sample and 123 participants in the validation sample. The estimation method used is the Maximum Likelihood (ML) method as suggested by Vandewaetere and Desmet (2009). ML estimation tends to be more stable and shows higher accuracy in terms of model fit, compared to other estimators (that would be suggested to be used in case of non-normality) such as generalized- and weighted least squares methods.

Descriptives

The descriptives of the items for participants from the calibration and validation sample are presented in Table 6. Item 2 was deleted as suggested by the EFA. Again, most scores range between 3 and 4 except for item 4 ('I will be playing a game rather than learning proportional reasoning') in both samples. Scoring higher than 3 on all items again indicates a positive trend of students' perceptions towards the goals and usefulness of educational games (i.e. learning proportional reasoning can be enhanced by playing an educational game).

Table 6

Descriptive Statistics (M and SD) for the Calibration and Validation Sample

	Item	Calibration (n = 119)		Validation (n = 123)	
		M	SD	M	SD
Item 1	I believe playing this game could be of some value to me.	3,77	1,26	3,56	1,36
Item 3	I think playing this game is important to do because it can help me to learn proportional reasoning.	3,83	1,19	3,37	1,29
Item 4	I will be playing a game rather than learning proportional reasoning.	4,32	1,45	4,39	1,53
Item 5	I think playing this game could help me to learn proportional reasoning.	3,80	1,17	3,41	1,26

Table 6 continued

	Item	Calibration (n = 119)		Validation (n = 123)	
Item 6	I will characterize my experience with the game as 'playing' rather than learning.	3,54	1,28	3,12	1,23
Item 7	When I will use the game, it will feel like playing rather than learning.	3,67	1,22	3,31	1,25
Item 8	I would be willing to play this game because it has some value to me to learn proportional reasoning.	3,43	1,41	3,13	1,30
Item 9	I think that playing this game is useful for learning proportional reasoning.	3,76	1,31	3,46	1,27

Confirmatory factor analysis (CFA)

First, the data was assessed for suitability for factor analysis for both groups (calibration and validation sample). The Kaiser-Meyer-Olkin (KMO) measure verified the sampling adequacy for the analysis for the calibration sample, as well as for the validation sample. The $KMO_{calibration}$ = .78 is good and the $KMO_{validation}$ = .82 is very good (Field, 2009; Hutcheson & Sofroniou, 1999). Barlett's test of sphericity $\lambda^2_{calibration}$ (28) = 414.77, $p < .001$ and $\lambda^2_{validation}$ (28) = 477.60, $p < .001$ indicated that correlations between items were sufficiently large and that there is some scope for reducing the number of dimensions (Field, 2009). Therefore we deemed CFA appropriate. Based on the outcomes of the EFA, we hypothesised a two –factor model to be confirmed in the measurement portion of the model. Several fit indices with their cut-off criteria were used to assess this model of fit, following Hooper, Coughlan, and Mullen (2008), Kline (2011), Schreiber, Nora, Stage, Barlow, and King (2006), and Vandewaetere and Desmet (2009). The theoretical model is presented in Figure 1.

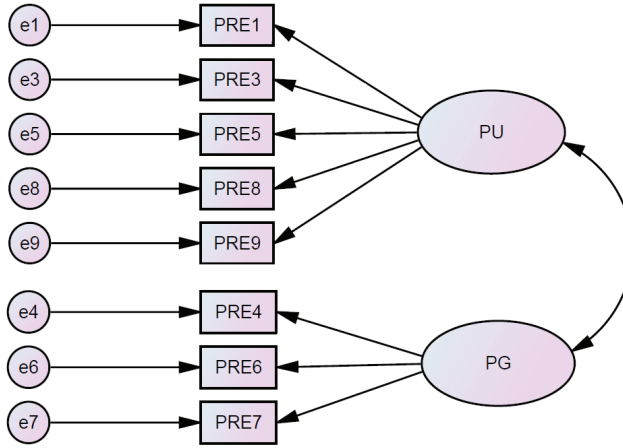

Figure 1. Visualization of the theoretical construct of the GPS with the two subscales PU and PG. The numbering of items (called PRE) is in accordance with Table 2. Associated with each observed variable (the PRE items), a measurement error (e) is represented.

The goodness-of-fit indices (Table 7) indicate a good fit between the model and the observed data for the calibration sample since all the indices meet the more recently set severe cut-off points (Browne & Cudeck, 1992; Byrne, 2001; Hooper et al., 2008; Hu & Bentler, 1999; Kline, 2011; Schreiber et al., 2006). No post-hoc modifications were indicated from the analysis because the initial model fits well and it is assumed to be unnecessary to modify a good model to achieve even better fit because these modifications may simply be fitting small idiosyncratic characteristics of the sample (MacCallum, Roznowski, & Necowitz 1992).

Table 7

GPS: Goodness-of-fit indices for the two-factor structure tested with CFA

	n	χ^2 (*df*)	p	CMIN/DF	RMSEA	GFI	AGFI	CFI
Calibration sample	114	25.939 (19)	0.132	1.365	0.057	0.952	0.909	0.983
Validation sample	128	23.211 (19)	0.228	1.222	0.042	0.958	0.920	0.991

Using the validation sample, the construct structure of the model was confirmed. Table 7 also presents the goodness-of-fit indices for the two-factor solution of the validation sample. Given these good values of the indices (Browne & Cudeck, 1992; Byrne, 2001; Hooper et al., 2008; Hu & Bentler, 1999; Kline, 2011; Schreiber et al., 2006) we are allowed to conclude for a good fit and no modification was necessary. Figure 2 and 3 display the standardized factor loadings of the two-factor model for the calibration and the validation sample.

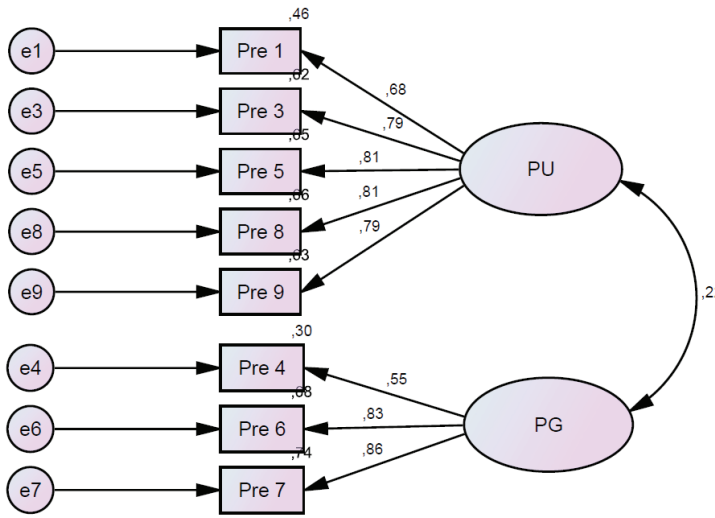

Figure 2. Calibration sample: Standardized factor loadings of the two-factor model as obtained with CFA. Numbering of items are in accordance with the numbering in Table 2.

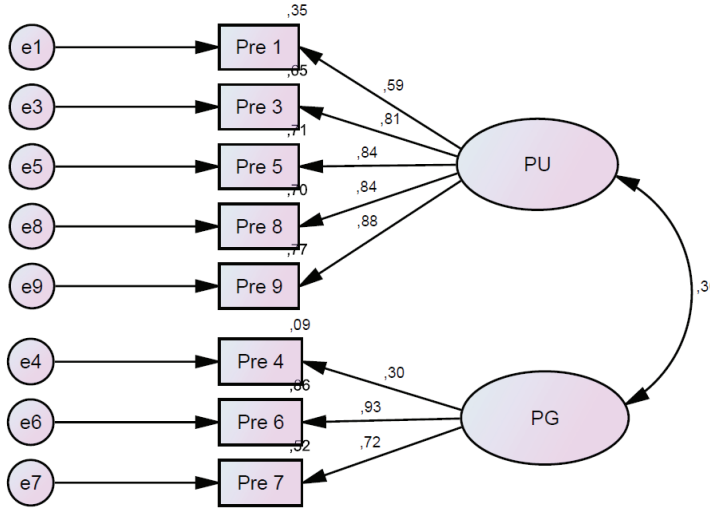

Figure 3. Validation sample: Standardized factor loadings of the two-factor model as obtained with CFA. Numbering of items are in accordance with the numbering in Table 2.

The intercorrelations are moderate to good. The PG subscale is positively correlated to the PU subscale for the calibration sample (r = .22) as well as for the validation sample (r = .30). This positive correlation between both subscales refers to the internal consistency of the questionnaire. The loadings of the items associated with both subscales, for both samples, are varying between acceptable (lowest loading = .30) and high (highest loading = .90).

Discussion

In the first part of this study, we identified and corroborated the two underlying factors of our definition of students' perceptions, being PU and PG. The results indicate that both factors are represented in the underlying structure of the GPS. At first sight, the internal consistency of the PG-subscale seemed unsatisfying. The second part of the study served to test the theoretical framework and provides additional validation. As the reliability results suggest, the psychometric properties maintained and the scale remains internally consistent for the complete scale and for the PU-subscale. The

PG-subscale is quite reliable, but does not always meet the severe suggested values of alpha according to Cortina (1993).

The CFA results show that the theoretical 2-factor model was confirmed and thus seemed satisfactory. The GPS seems to be a valid operationalization for measuring vocational students perceptions of a GBLE. Because of interest in the reliability and validity of the GPS with other target groups and different learning environments, a second study was initiated.

PART 2: MULTIPLE GROUP CFA: GENERALISING THE VALIDITY

For further validation of the GPS across different participants, type and use of GBL, we merged data from three different studies in which we administered the GPS questionnaire. Because in our future studies we are interested in determining what factors influence student's perceptions, and thus in turn, influence GBL processes, we also measured students' perceptions after students' interaction with a GBLE. So in the following part, pre- and post-measurements will be taken into account. Next to this, we want also want to test for the invariance of both the items and the factorial structure across groups (Byrne, 2001), i.e. for different education levels, for different GBLEs and for a different play time opportunity for participants.

Method

As just mentioned, data from three different studies was merged. All studies had a similar pre-posttest design and participants only differed with respect to their educational level (primary school pupils or vocational secondary school students), the game environment they played with (MonkeyTales[2] or Zeldenrust) and the amount of play time participants had (some could choose by their own how long they wanted to play, others had less than one hour play-time or more than one hours play time; see Table 8). These are the three group subdivisions we will focus on during the multiple group CFA. The 8-item GPS was assessed on paper before and after gameplay.

Table 8
Overview of studies: subdivision in three groups for multiple group CFA

	n	Educational Level	Game Environment	Duration Play time
Study 1	101	Primary School (7-9 year)	Monkey Tales (commercial GBLE)	Pupils decide how long they played
Study 2	135	Vocational Secondary School (14-17 year)	Zeldenrust (research GBLE)	1 hour
Study 3	92	Vocational Secondary School (14-17 year)	Monkey Tales (commercial GBLE)	2 hours

Results

Of the 328 participants, 277 completely filled in the GPS questionnaire before they started playing the game (8 items on a 6-point Likert-scale) and 212 participants completely filled in the GPS after the gameplay[3]. AMOS software was used to perform the CFA and multiple group CFA.

Data preparation

Outliers were again detected following the same procedure according to Kline (2011) and as described above. Eight observations were considered as outliers (four for the pre-GPS and post-GPS). In line with the previous study, ML estimation was used to analyse the data.

Multiple group CFA (see Footnote 4)

Prior to testing for invariance across multigroup samples, it is customary to first consider a baseline model, estimated for each group separately (Byrne, 2001). Based on the findings in our previous analyses, item two was excluded and the 8-item GPS model was tested separately for each group (6 in total) as baseline model. Findings were consistent, for pre- and post-measurement of the GPS, in always revealing correlated errors between item 9 ('I think that playing this game is useful for learning proportional reasoning') and item 8 ('I would be willing to play this game because it has some value to me to learn proportional reasoning'). This correlation indicates overlap between this pair of items (Byrne, 2001). When looking more closely to the items, both are indeed explicitly focused on the learning value of the game. Although the other items of this subscale also measure the use-

fulness of the game, item 8 and 9 are both more directed to the prospective perceived usefulness of the GBLE. Because of this, a correlation between both errors will be added to the model. This results in the baseline model that is identically specified for each of the groups (see Figure 4).

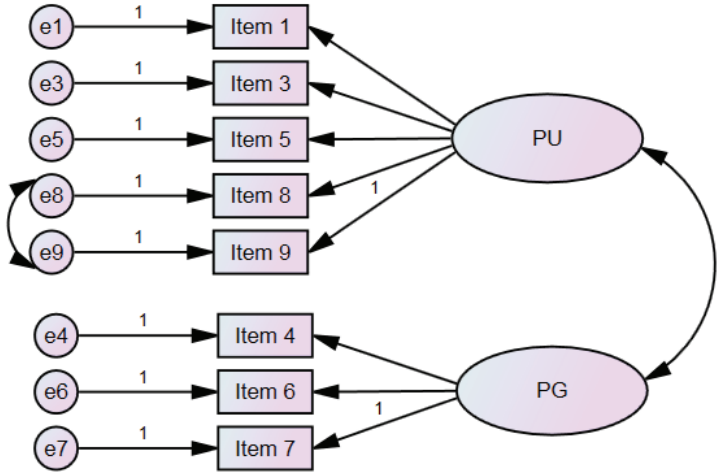

Figure 4. Baseline model of 8-item GPS for all groups tested, as well for pre-, as for post- measurement.

In the following parts, we test for invariance across the three different group divisions. For the process of determining nonequivalence of measurement and structural parameters across groups we will use the procedures as described by Byrne (2001) which involves the testing of a series of increasingly restrictive hypotheses. The orderly sequence (as described below) of the analytic steps is both necessary and strongly recommended (Byrne, 2001).

Education level

As a first step in testing for invariance across pupils in primary versus students in secondary school, we test for the validity of the GPS structure (pre- and post-measurement) as best represented by the hypothesized two-factor structure as shown in Figure 4. Parameters are estimated for both groups at the same time and provides the baseline value against which all subsequent models are compared. In Table 9, the goodness-of-fit statistics for the structure of the pre-GPS ($\chi^2_{(36)} = 42{,}52$; Model 1a) and post-GPS ($\chi^2_{(36)} = 59{,}12$; Model 1b) which provide the baseline models are reported.

The modification indices (CFI_{pre} =0.992, $RMSEA_{pre}$ = 0.026, CFI_{post} = 0.974, $RMSEA_{post}$ = 0.055) indicate that the hypothesized two-factor model is well-fitting across the two education levels. Although the indices of the GPS-post show slightly lower values than the indices of the pre-measurement, they still meet the severe cut-off points (Browne & Cudeck, 1992; Byrne, 2001; Hu & Bentler, 1999).

After establishing the good fit of this model, we proceed in testing the invariance of factorial measurement and structure across groups by placing constraints on particular parameters by specifying them as invariant (i.e., equivalent) (Byrne, 2001). As suggested by Byrne (2001), we start with a test for the possibility that a fully constrained model (all factor loadings, all factor variances, all factor covariance and the error variance are constrained equal) is invariant across groups. This model is excessively stringent, but it seems prudent to ascertain whether the error covariance holds across education level (Byrne, 2001). The comparisons of these fully constrained models (models 2a and 2b) with their baseline models ($\chi^2_{pre(62)}$ = 169.83, $\Delta\chi^2_{pre(26)}$ = 127.31, $p < .001$; $\chi^2_{post(62)}$ = 224.81, $\Delta\chi^2_{post(26)}$ = 165.69, $p < .001$) indicate that some equality constraints do not hold across the two education levels for the pre- as well as for the post-measurement. In the following part the equality of sets of parameters is tested to gain more insight.

In a third step, we test for the equivalence of all factor loadings across the two groups. As indicated in Table 9 all factor loadings are equivalent across pupils and secondary school students, for the pre- as for the post-measurement, as the chi-square difference between both models is not statistically significant (p_{pre} = .207, p_{post} = .079). This indicates that we can be confident that the GPS serves as a valid and reliable instrument for measuring pupils' and students' perceptions before and after gameplay.

In a following step the equality of the structural parameters, e.g., the invariance of factor variances and covariance across both groups is tested. For the pre-measurement, results yielded a statistically significant chi-square value ($\chi^2_{pre(45)}$ = 61.48, $\Delta\chi^2_{pre(9)}$ = 18.96, p = .026). The next step according to Byrne (2001) is to determine which variances are contributing to this inequality. As indicated in Table 9, the tests (model 5a-9a) revealed all covariances and variances to be equal across pupils and secondary school students. When looking more closely to the significant difference found in model 4a, the covariance and variance of PU subscale were comparable. We did found a striking higher variance for the primary school pupils concerning the PG subscales. This indicates a greater heterogeneity between the pupils concerning the PG of a GBLE than in the secondary school group. This finding is in line with the constitution of the classes in primary (all children

together) and secondary school (more differentiated, i.e., vocational track students).

For the post-measurement, the same test for invariance of factor variance and covariance across both groups was done. A statistically significant chi-square value was found ($\chi^2_{post(45)} = 94.35$, $\Delta\chi^2_{post(9)} = 35.23$, $p < .037$). Again, we determined which variances are contributing to this inequality. Models 5b-9b (see Table 9) revealed all covariances and variances to be equal across pupils and secondary school students, except for two combination; model 8b and 9b. This indicates that both groups (pupils and secondary school students) showed significant differences with respect to variance of PG and PU and with respect to the covariance between subscales. The difference in the normalized version of the covariance the correlation coefficient, shows that the strength of the linear relation between both subscales differs between the groups (i.e. higher for the primary school children) after playing an educational game. Additionally, for both models (model 8b and 9b) the variance of PG and PU was higher for the secondary school children, indicating higher heterogeneity after playing the game, while this was not the case before they played the games.

Table 9

GPS: Goodness-of-fit statistics for tests of invariance – Education Level: a summary

	Model Description	χ^2	df	$\Delta\chi^2$	Δdf	Statistical significance
Pre	Model 1a (Hypothesized model)	42.52	36	-	-	-
	Model 2a (factor loadings, variances and covariances constrained equal)	169.83	62	127.31	26	p < .001
	Model 3a (factor loadings constrained equal)	50.97	42	8.45	6	p = .207
	Model 4a (Model 3a with all variances and covariance constrained equal)	61.48	45	18.96	9	p = .026
	Model 5a (Model 3a with variance of PU constrained equal)	51.94	43	9.42	7	p = .223
	Model 6a (Model 3a with variance of PG constrained equal)	52.72	43	10.20	7	p = .18

Table 9 continued

	Model Description	χ^2	df	$\Delta\chi^2$	Δdf	Statistical significance
	Model 7a (Model 3a with variance of PG and PU constrained equal)	56.53	44	11.01	8	p = .201
	Model 8a (Model 5a with equal covariance)	55.64	44	13.12	8	p = .108
	Model 9a (Model 6a with equal covariance)	57.31	44	14.791	8	p = .063
Post	Model 1b (Hypothesized model)	59.12	36	-	-	-
	Model 2b (factor loadings, variances and covariances constrained equal)	224.81	62	165.69	26	p < .001
	Model 3b (factor loadings constrained equal)	70.43	42	11.31	6	p = .079
	Model 4b (Model 3b with all variances and covariance constrained equal)	94.35	45	35.23	9	p = .037
	Model 5b (Model 3b with variance of PU constrained equal)	71.29	43	12.17	7	p = .095
	Model 6b (Model 3b with variance of PG constrained equal)	72.03	43	12.91	7	p = .074
	Model 7b (Model 3b with variance of PG and PU constrained equal)	72.70	44	13.58	8	p = .093
	Model 8b (Model 5b with equal covariance)	84.59	44	25.47	8	p < .001
	Model 9b (Model 6b with equal covariance)	86.49	44	23.37	8	p < .001

Game environment

The same analysis and procedure as for the educational level groups were performed. Here we want to determine if measuring students perceptions with the GPS (with the two subscale scores) is equivalent across students who played with the different GBLE's. Because this implies we want

to measure participants' experience with the GBLE, only post-measurement is taken into account.

The baseline value is reported in Table 10 ($c^2_{(36)}$ = 60.11; model 10). Again the modification indices (CFI = 0.976 and RMSEA = 0.057) indicate that the hypothesized model is well-fitting across the two GBLEs (Browne & Cudeck, 1992; Byrne, 2001; Hu & Bentler, 1999). Subsequently, the comparison of the fully constrained model (model 11) with the baseline model yields a chi-square difference ($\Delta\chi^2$) value of 75.11 with 26 *df,* which is statistically significant ($p < .001$). This means that some equality constraints do not hold across the two GBLEs. To gain more insight, the equivalence of all factor loadings across the two groups (model 12) is tested. Because the chi-square difference is not statistically significant ($p = .469$) we can be confident that measuring students perceptions with the GPS after they have played the games is operating in the same way for the two GBLEs. In a next step, the invariance of factor variances and covariance across both groups are tested and yields a statistically significant chi-square value ($\Delta\chi^2_{post(9)} = 25.83$, $p = .002$). The tests (models 14-18; see Table 10) reveal all covariances and variances to be equal across both GBLEs except for the models 16, 17 and 18. More specific, the group that played with the GBLE that was developed for research purposes (Zeldenrust) showed higher correlation between PU and PG. Additionally, this group shows lower variance for both subscales, indicating less heterogeneity between the students concerning their PU and PG of the GBLE.

Table 10

GPS: Goodness-of-fit statistics for tests of invariance – Game Environment: a summary

	Model Description	χ^2	df	$\Delta\chi^2$	Δdf	Statistical significance
Post	Model 10 (Hypothesized model)	60.11	36	-	-	-
	Model 11 (factor loadings, variances and covariances constrained equal)	135.22	62	75.11	26	p < .001
	Model 12 (factor loadings constrained equal)	65.72	42	5.61	6	p = .469
	Model 13 (Model 12 with all variances and covariance constrained equal)	85.94	45	25.83	9	p = .002

Table 10 continued

	Model Description	χ^2	df	$\Delta\chi^2$	Δdf	Statistical significance
	Model 14 (Model 12 with variance of PU constrained equal)	73.57	43	13.46	7	p = .062
	Model 15 (Model 12 with variance of PG constrained equal)	70.94	43	10.83	7	p = .146
	Model 16 (Model 12 with variance of PG and PU constrained equal)	75.88	44	15.77	8	p = .046
	Model 17 (Model 14 with equal covariance)	76.10	44	15.99	8	p = .043
	Model 18 (Model 15 with equal covariance)	72.33	44	12.22	8	p = .014

Game play time

To investigate whether the GPS also holds for short-term and long-term interventions, the same analyses and procedure are done with data gathered from students who played only one hour (or less) with a GBLE, and students who played longer than one hour. Also for these analyses, only post-measurement is taken into account, since the model structure for the pre-measurement was equal for both short- and long-term gameplay.

The baseline value is $c^2_{(36)} = 50.53$ (model 19; see Table 11) and the CFI and RMSEA values of 0.984 and 0.045 respectively show that the hypothesized model is very well-fitting across the two groups (Browne & Cudeck, 1992; Byrne, 2001; Hu & Bentler, 1999). The comparison of the fully constrained model (model 20) and model 19 yields a statistically significant chi-square difference ($\Delta\chi^2_{(26)} = 54.21$, $p = .001$) and thus indicates that some equality constraints do not hold across both groups. The test for the equivalence of all factor loadings across both groups (model 21) is not statistically significant ($p = .580$; Table 11). Again we can be confident that the GPS is operating in the same way for both groups, the short- and long-term gameplay group, after they have played the game. Finally, factor variances and covariances across both groups are tested. The difference in chi-square value of 25.83 with 9 *df* between this model 22 and model 19 is statistically significant ($p = .002$). Consequently we determine which variances are contributing to this in-equality. As indicated in Table 11, the tests (models

23-27) revealed all covariances and variances to be equal across both groups except for models 26 and 27. Significant difference is found with respect to the variance of PU and PG (with a higher variance for PG and PU for the group that played more than one hour). This indicates that students who played longer with the game, showed more heterogeneity in their PU and PG of the GBLE. This in combination with a higher covariance for the group that played less than one hour, we can conclude that the longer students play with a game, the more diverse their PU and PG of the GBLE.

Table 11

GPS: Goodness-of-fit statistics for tests of invariance – duration: a summary

	Model Description	χ^2	df	$\Delta\chi^2$	Δdf	Statistical significance
Post	Model 19 (Hypothesized model)	50.53	36	-	-	-
	Model 20 (factor loadings, variances and covariances constrained equal)	104.74	62	54.21	26	p = .001
	Model 21 (factor loadings constrained equal)	55.25	42	4.72	6	p = .580
	Model 22 (Model 21 with all variances and covariance constrained equal)	73.34	45	22.81	9	p = .007
	Model 23 (Model 21 with variance of PU constrained equal)	63.95	43	13.42	7	p = .062
	Model 24 (Model 21 with variance of PG constrained equal)	58.54	43	8.01	7	p = .332
	Model 25 (Model 21 with variance of PG and PU constrained equal)	65.01	44	14.48	8	p = .070
	Model 26 (Model 23 with equal covariance)	66.65	44	16.12	8	p = .041
	Model 27 (Model 24 with equal covariance)	59.09	44	8.56	8	p = .038

Discussion

In this second part, the GPS was examined with different samples to represent different populations. The results indicate that the theoretical hypotheses of the two-factor model is confirmed for the factor loadings for students of primary and secondary education, for the two different GBLEs and for different durations of gameplay. These findings imply that we can be confident that in measuring students' and pupils' perceptions with the GPS, we are measuring the same construct regardless the amount of play time, the GBLE that is played in and the educational level of the participants. The stable structure of the GPS, together with its good reliability, makes this scale suitable for research purposes in different target groups without violating the psychometric properties. Testing the equivalence of the model structure across groups revealed some slight differences with respect to the variance of the subscales and the covariance between the subscales for some groups.

When researchers want to measure pupils' and students' perceptions of GBLE's before they got the chance to play the game (prospective), but also after students experiences the GBLE (retrospective), the GPS seems to be a reliable instrument.

SUMMARY AND CONCLUDING DISCUSSION

For the development of the GPS, research on students' perceptions was consulted. Subscales from the IMI and the PES were adjusted in order to fit the requirements for this scale: providing a measurement of students' and pupils' perceived usefulness and goals of GBLEs. We defined perceptions as (1) students' expectations about the goals of the environment and (2) the degree to which a student believes that using a GBL enhances his or her performance on the domain that is focused on in the GBLE. In a first study, the EFA revealed a meaningful two-factor solution, which reflects the two dimensions that constitute the notion of game perceptions. Further, we used a CFA to validate and confirm the model. Cross-validation was repeated in second part of this article in which a large sample was used to perform multi group CFAs. We compared groups with different education levels, different GBLEs and different amounts of play time. The results showed that GPS can be used in different target groups when researchers want to measure students' perceptions of GBLE's for both pro- and retrospective purposes. The observed measures were found to be operating equivalently among the game

perception components (PU and PG). However, significant differences in structural relations with respect to the covariance and the variance of the PU and PG subscale were found for all the groups. These findings indicate that the educational level of students, the GBLE they play with and the amount of time students play with a GBLE influences the strength of the linear relation between their PU and PG of GBLE, as well as the diversity of the perspective of students' perceptions.

If this is kept in mind (i.e. constrained equal) when the GPS is used for research purposes the GPS can be used for different target groups, for different GBLEs and for different amounts of play-time without violating the psychometric properties. The GPS seems to be a reliable and stable way to measure students' perceptions about the goal and usefulness of a GBLE before and after they are confronted with the learning environment. Based on the study of Salomon (1984) and the mediational paradigm of Winne (1987) one might expect that students who perceive games as a leisure time activity (no learning goal) will invest less mental effort to process the information specific for learning, which may result in fewer learning gains. By contrast, learners who perceive the environment as a learning environment (and expect a learning goal) may invest more mental effort and have higher learning gains. The effectiveness of the GBLE may thus be largely affected by students' perception about the environment. With the development of the GPS and the abovementioned studies, we've set some first and preliminary steps towards the positioning of perceptions into a broader framework of factors that influence the GBL process. Hopefully, the scale will spark a surge of interest into investigating how students' perceptions of a GBLE influence the implementation of GBLE's in educational settings.

References

Becker, K. (2007). Pedagogy in commercial video games. In D. G Gibson, C. A. Aldrich, & M. Prensky (Eds.), *Games and simulations in online learning: Research and development frameworks.* Hershey, PA: Information Science Publishing. doi: 10.4018/978-1-59904-304-3.ch002

Brown, J. D. (2009). Choosing the right type of rotation in PCA and EFA. *JALT Testing & Evaluation SIG Newsletter, 13*(3), 20-25.

Browne, W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*, 230-258. doi: 10.1177/0049124192021002005

Byrne, B. M. (2001). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming* (2nd ed.). New York, NY: Taylor and Francis Group.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education, 59*, 661-686. doi: 10.1016/j.compedu.2012.03.004

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104. doi: 10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. doi: 10.1007/BF02310555

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*, 319-340. doi: 10.2307/249008

Deci, E. L., Eghrari, H., Patrick, B. C., & Leone, D. R. (1994). Facilitating internalization: The self-determination theory perspective. *Journal of Personality, 62*, 119-142. doi: 10.1111/j.1467-6494.1994.tb00797.x

de Freitas, S. (2006). *Learning in immersive worlds: A review of game-based learning*. Technical report prepared for Joint Information System Committee (JISC): London.

Elen, J., & Lowyck, J. (2000). Instructional metacognitive knowledge: A qualitative study on conceptions of Frenchmen about instruction. *Journal of Curriculum Studies, 32,* 421-444. doi: 10.1080/002202700182637

Entwistle, N. (1991). Approaches to learning and perceptions of the learning environments: Introduction to the special issue. *Higher Education, 22,* 201-204.

Field, A. P. (2009). *Discovering statistics using SPSS (3th edition).* London: Sage.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming, 33*, 441-467. doi: 10.1177/1046878102238607

Hays, R. T. (2005). *The effectiveness of instructional games: A literature review and discussion* (Technical Report No. 2005-004). Naval Air Warfare Center Training Systems Division, Orlando, FL.

Hayes, E., & Games, I. (2008). Learning through game design: a review of current software and research. *Games and Culture, 3*, 309-332.

Hooper, D., Coughlan, J. & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods, 6*, 53-60.

Hu, L-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: A multidisciplinary Journal, 6*, 1-55. doi: 10.1080/10705519909540118

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist.* London: Sage.

Ke, F. (2008). A case study of computer gaming for math: Engaged learning from gameplay? *Computers & Education, 51,* 1609-1620. doi: 10.1016/j.compedu.2008.03.003

Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. In R. Ferdig (Ed.), *Handbook of research on effective electronic gaming in Education* (pp. 1-32). doi:10.4018/978-1-59904-808-6.ch001

Kebritchi, M. (2008). *Effects of a computer game on mathematics achievement and class motivation: An experimental study.* Unpublished doctoral dissertation, University of Central Florida.

Kirriemuir, J., & McFarlane, A. (2004). *Literature Review in Games and Learning* (No. 8). Bristol, UK: Nesta Futurelabs.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press. doi: 10.1111/insr.12011_25

Liu, T. Y., & Chu, Y. L. (2010). Using ubiquitous games in an English listening and speaking course: Impact on learning outcomes and motivation. *Computers & Education*, *55*, 630-643. doi: 10.1016/j.compedu.2010.02.023

Lowyck, J., Elen, J., & Clarebout, G. (2004). Instructional conceptions: Analyses from an instructional design perspective. *International Journal of Educational Research, 41,* 429-444. doi: 10.1016/j.ijer.2005.08.010

MacCallum, R. C. Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis the problem of capitalization on chance. *Psychological Bulletin, 111*, 490-504.

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research,3*(1), 97-110.

McAuley, E., Duncan, T., & Tammen, V. V. (1987). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport, 60,* 48-58. doi: 10.1080/02701367.1989.10607413

O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *The Curriculum Journal, 16*, 455-474. doi: 10.1080/09585170500384529

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: impact on educational effectiveness and school motivation. *Computer & Education, 52*, 1-12. doi: 10.1016/j.compedu.2008.06.004

Pavlas, D., Jentsch, F., Salas, E., Fiore, S. M., & Sims, V. (2012). The play experience scale: Development and validation of a measure of play. *Human Factors: The Journal of The Human Factors and Ergonomics Society,54*, 215-225. doi: 10.1177/0018720811434513

Ritzhaupt, A., Higgins, H., & Allred, B. (2011). Effects of modern education game play on attitudes towards mathematics, mathematics self-efficacy, and mathematics achievement. *Journal of Interactive Learning Research, 22*, 277-297. Chesapeake, VA: AACE.

Salomon, G. (1984). Television is "easy" and print is "though": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology, 76*, 647-658.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King J. (2006). Reporting structural equation modeling and confirmation factor analyses results: A review. *The Journal of Educational Research, 99*, 323-338. doi: 10.3200/JOER.99.6.323-338

Shih, C.-L., & Chuang, H.-H. (2013). The development and validation of an instrument for assessing college students‘ perceptions of faculty knowledge in technology-supported class environments. *Computers & Education, 63*, 109-118. doi: 10.1016/j.compedu.2012.11.021

Shuell, T. J., & Farber, S. L. (2001). Students' perceptions of technology use in college courses. *Journal of Educational Computing Research, 24*, 119-138. doi: 10.2190/YWPN-H3DP-15LQ-QNK8

Struyven, K., Dochy, F., Janssens, S., & Gielen, S. (2008). Students' experiences with contrasting learning environments: The added value of students' perceptions. *Learning Environments Research, 11,* 83-109. doi: 10.1007/s10984-008-9041-8

Vogel, J. J., Greenwood-Ericksen, A., Cannon-Bowers, J., & Bowers, C. A. (2006). Using virtual reality with and without gaming attributes for academic achievement. *Journal of Research on Technology in Education, 39*, 105-119.

Vandercruysse, S., ter Vrugte, J., Wouters, P., Clarebout, G., Elen, J. (2012). Development of a mathematical game-based learning environment for vocational students. EARLI Sig 6&7. Bari (Italy), 11-13 September 2012.

Vandercruysse, S., Vandewaetere, M., Clarebout, G. (2012). Game based learning: A review on the effectiveness of educational games. In M. Cruz-Cunha (Eds.), *Handbook of Research on Serious Games as Educational, Business, and Research Tools* (pp. 628-647). Hershey, PA: IGI Global. doi: 10.4018/978-1-4666-0149-9.ch032

Vandewaetere, M., & Desmet, P. (2009). Introducing psychometrical validation of questionnaires in CALL research: The case of measuring attitude towards CALL. *Computer assisted language learning: an international journal, 22*, 349-380. doi: 10.1080/09588220903186547

Vandewaetere, M., Vandercruysse, S., Clarebout, G. (2012). Learners' perceptions and illusions of adaptivity. *Educational Technology Research and Development, 60*, 307-324. doi: 10.1007/s11423-011-9225-2

Winne, P. H. (1987). Why process-product research cannot explain process-product finding and a proposed remedy: the cognitive mediational paradigm. *Teaching and Teacher Education, 3*, 333-356. doi: 10.1016/0742-051X(87)90025-4

Wouters, P., van der Spek, D., & van Oostendorp, H. (2009). Current practices in serious game research: A review from a learning outcomes perspective. In T. M. Connolly, M. Stansfield, & L. Boyle (Eds.), *Games based learning advancements for multisensory human computer interfaces: Techniques and effective practices* (pp. 232-255). Hershey, PA: IGI Global. doi: 10.4018/978-1-60566-360-9.ch014

Footnotes

[1]AMOS software was used to perform the CFA.

[2]MonkeyTales concerns a 3-D existing educational game that students play individually. The game blends fun and learning, based on proved didactic methods. As part of the overall gameplay and in order to be able to advance in the game, students have to compete with in-game characters and to play mathematical mini-games which are related to comparing, adding and multiplying fractions.

[3]This attrition can be ascribed to different aspects. First, the questionnaire needed to be filled in completely to be able to analyse the data. Also students who filled in all but one item, were removed from data. Next to this, in the first study, some participants dropped out (n = 27) because playing the game was not obligatory. Third, during the studies we experienced technical problems (some participants did not have the chance to properly play the game) and also the spacing of the studies over several math-course hours (a considerable amount of participants did not attend all four hours) influenced this drop-out rate.

[4]AMOS software was used to perform the CFA and multiple group CFA.