Gamification in assessment

The effect of gamification on the response behaviour of students in an English vocabulary multiple-choice assessment

KEYWORDS: gamification, game-based assessment, games in education, assessment, validity

Researcher: Maaike Grobben

University of Twente

MSc Educational Science and Technology

Examination committee

First supervisor: prof. dr. ir. B.P. Veldkamp

Second supervisor: dr. J.W. Luyten

External organisation: Centraal Instituut voor Toetsontwikkeling (CITO)

External supervisor: R. Noordhof

Second external supervisor: dr. R.C.W. Feskens

Acknowledgement

First, I would like to thank my thesis supervisor prof. dr. ir. B.P. Veldkamp of the department Research Methodology, Measurement and Data Analysis of the faculty of Behavioural, Management and Social Sciences at the University of Twente. I would like to thank him for his support from the beginning to the end of writing my thesis.

I would also like to thank stichting Cito for facilitating my internship. A special thanks goes out to Romy Noordhof, R&D-specialist at Citolab, who guided me through the whole process of writing my thesis. Also, I would like to thank Remco Feskens, Senior Research Scientist at Cito for his help with the analysis of the data. Last, I would like to thank the ICT-team of Citolab, and especially Patrick de Klein, for creating the game I used in this master thesis.

Last, I would also like to acknowledge dr. J.W. Luyten of the department Research Methodology Measurement and Data Analysis of the faculty of Behavioural, Management and Social Sciences at the University of Twente as the second reader of this thesis.

Abstract

Currently, gamification generates a lot of attention in education as a method to increase student motivation and engagement. Since students are often nervous or stressed before examination, gamification also appears to be a promising strategy in educational assessment. Previous research regarding game-based assessment focused on effects and how to put it into practice, but it is unclear to what extent gamification affects the response behaviour of students and with that, the validity of the test. Therefore, this research has been carried out to find out to what extent gamification of an English vocabulary multiple-choice assessment affects the response behaviour of students. A traditional computer-based condition is compared to three game-based conditions with different game interactions: clicking, swiping, and shooting. In the design process, content validity is ensured by conducting a pilot test and making informed decisions. The results show that gamification does not affect the test scores of students, and therefore gamification seems not to affect the construct validity of the test. Research on student characteristics show that overall boys score higher than girls, but there is no significant difference within the conditions. Even though the duration of the test is longer, students in the game conditions report to be more motivated and appreciate this way of testing more than a traditional test. Also, student motivation is positively related with test score. An unexpected result is that a positive relationship has been found between game experience and test score in the traditional test and the clicking game, while it was expected that these conditions require limited or no game skills. Generally, the results indicate that gamification does not influence the test scores of students and therefore this could be a valid method to increase student motivation in assessment contexts. However, further research is needed to explain unexpected findings before game-based assessment will be used for high stake decisions in education.

Table of contents

Introduction

Students in the Netherlands are less motivated to learn and achieve their learning goals, compared to other wealthy countries (Inspectie van het Onderwijs, 2019). This is alarming, since the results of Dutch students have also been slightly decreasing for the past 20 years. On top of that, most children feel nervous or stressed before examination, which negatively influences their working memory (Aydin, 2019; Lewis, Nikolova, Chang, & Weekes, 2008). Therefore, it is worth considering new assessment methods to increase motivation and decrease stress and nervousness.

A method that increases student motivation and engagement is gamification; applying game-related elements to nongame contexts (Dominiguez et al., 2013; Kapp, 2012; Prince, 2013). Gamification has currently generated attention across a range of contexts, such as education, human resource management (HRM) and marketing. In education, it is mainly used for instructional purposes (Buckley & Doyle, 2016). However, it is also a very promising strategy for assessment purposes, because in previous research a negative relationship has been found between stress and motivation (Park et al., 2012). Since gamification increases student motivation and engagement, it could be an excellent method to decrease examination stress and increase students' excitement for testing.

Next to the positive effects of gamification, there are also some downsides to this concept. Gamification is seen as an exciting, promising trend, but the disadvantages of a game-based approach in educational assessment are investigated inadequately. For example, students could be distracted by the game or the measured ability of the test could (unconsciously) be: who has the best gaming skills? It could be that a student who is not familiar with gaming, performs worse on game-based assessment than on a traditional test. In other words: a gamified assessment may affect the test scores and other response behaviours of the students, which might affect the validity of the test.

Some research has been carried out on the validity of game-based personality assessment. For example, Ventura and Shute (2013) found that a valid assessment of persistence can be achieved in a video game and Denden, Tlili, Essalmi, and Jemni (2018) found that gaming behaviours can model learners' personality. However, before gamification can be implemented in education to make testing more fun and attractive, it is essential to find out to what extent this method influences the response behaviour of students, and especially the validity of students' test scores. This is important, because in classrooms important decisions and actions are based on information that is gathered through assessment, such as the instruction method and even an advice for a students' next level of education.

In this research, the response behaviour of students in three different game conditions will be compared to the response behaviour of students in a traditional computer-based test. The game interaction will be different in each game condition. The outcomes of this research can contribute to the decision if game-based assessment should be used on a large scale or not. In this research, an English vocabulary multiple-choice assessment will be used.

Theoretical background

Gamification

In gamification, game-related elements are applied to nongame contexts to engage people, motivate action, promote learning and solve problems (Kapp, 2012; Prince, 2013). The focus on these learning aspects, makes gamification very different from traditional games in which entertainment is often the only goal (Dominguez et al., 2013). The increasing interest in gamification arises from the idea that a game-setting can trigger natural behaviour and that this behaviour can be influenced by game-elements (McGonigal, 2011). In games, people tend to give different emotional responses, such as curiosity, frustration and joy. These responses are strengthened by including rewards and other motivators. Often, rewards are not directly related to the achieved goal, but they identify that a player has achieved a certain level of competence (Buckley & Doyle, 2016). As an example: a student receives coins as a reward in a language-learning game. With these coins, new game-elements can be bought. These coins indicate that this student has a certain level of competence, but the coins do not have a direct relationship with the learning goal.

Gamification is seen as a relatively new trend in education, but it has been around in a marketing context for a while (Prince, 2013). In most companies and shops, there is a save up system to get discount or a reward when you have spent a certain amount of money. This is similar to the game principle of reaching a certain level (of competence) in a game. In education and marketing, the basic principle is similar: gamification leads to increasing engagement and motivation (Prince, 2013). However, the overall goal is a little different. In marketing contexts, the goal is that customers feel motivated to come back, while in education the goal of gamification is to increase student motivation and thereby their achievement. Gamification is not only used to increase motivation; in HRM-contexts it is also used in the recruitment and selection process of employees (Armstrong, Ferrell, Collmus, & Landers, 2016). Game-based selection assessments can assess applicants' knowledge, skills, abilities and other characteristics, which could predict job performance (Lievens & De Soete, 2012). These aspects include important information during the selection process.

Effects of gamification

In recent research, several effects of gamification in education have been found. A general positive impact of gamification is that games promote learning and decrease the time of teaching. This positively influences many subjects for pupils of different age (Chandel, Dutta, Tekta, Dutta, & Gupta, 2015). Next to that, games in education have been found to increase student motivation and engagement (Dominguez et al., 2013; Papastergiou, 2009). Westrom and Shaban (1992) also found that games positively affect student motivation. However, as participants gained experience on a game, the game became less challenging and curiosity diminished, which reduced the motivation of the participants. Therefore, there should be a large variety of game elements in educational games. Another promising effect of games on learning has been found by Vogel et al. (2006). They found that learners who used

interactive games for learning had the greatest cognitive gains over learners provided with traditional classroom training.

Next to the positive effects on motivation, engagement and cognitive variables, educational games also encompass opportunities for self-assessment and self-evaluation (Suls, Martin, & Wheeler, 2002). Since people naturally involve themselves in social comparison, games take advantage of this natural process by using different rewarding game elements such as points, leader boards, and badges. By using these game-elements, self-assessment as well as self-evaluation can be enabled. The use of rewarding game elements ensures that students will be motivated to improve their own scores. Also, if leader boards are used, students will also be motivated to score higher than other classmates. These processes involve a large set of self-evaluative questions and self-motivations, which are important aspects in learning (Suls et al., 2002).

The effectiveness of gamification is not affected by gender or previous game experience (Nietfeld, Shores, & Hoffmann, 2014; Papastergiou, 2009). Girls perform at similar levels as boys in game-based learning environments, despite incoming disadvantages for perceived skill and prior gaming experience (Nietfeld et al., 2014). Previous game experience leads to advantages in early stages of game-based learning. Generally, boys have more game experience than girls, which means that boys have some advantages in the early stages of game-based learning. However, this gap is closed to the point of statistically nonsignificant differences by the end of the play (Nietfeld et al., 2014). This is in line with the results of Papastergiou (2009), who also found that there were no significant differences between boys and girls in game-based learning. Also, findings related to motivational variables, such as self-efficacy, goal orientation, and situational interest were not influenced by gender. These results indicate that using game-elements may be a promising method to increase motivation and excitement for testing. Therefore, the game-based elements used in learning environments are also gaining popularity in assessment contexts.

Types of assessment

Before focussing on game-based elements in assessment contexts, it is important to get understanding of what assessment is and which types of assessment are commonly used. According to Bennett (2011) assessment can be described as exploring domain understanding of students by designing tasks or asking questions. Generally, two assessment purposes are distinguished: summative assessment and formative assessment. Summative assessment is used at the end of the learning process and to assess the quality of learning (Black & Wiliam, 2009). Feedback is provided afterwards, and students do not have an opportunity to take this feedback into account to improve their learning on the subject that is assessed (Black & Wiliam, 2009). Also, summative feedback is often very limited: sometimes only a grade is provided. An advantage of summative assessment is that this method can be used to compare results with other classes, schools and sometimes even other countries. This makes summative assessment an important measurement in assessing and evaluating (the quality of) education.

The information obtained through formative assessment indicates what the current level of the students is. Then, it can be decided what students will need to develop themselves and achieve their learning goals (Hattie & Timperley, 2007). Formative assessment can be used by teachers, learners and their peers to make decisions about what the next steps of learning should be, for example the type of instruction, type of exercises and the type of questions a student asks (Black & Wiliam, 2009). This means that formative assessment is an important aspect in facilitating student learning (Panayiotis & James, 2013).

Apart from the different purposes of assessment, different types of assessment can be distinguished. The most common type of assessment is traditional assessment, in which different kind of (standardized) items are used: multiple-choice tests, true/false tests, short answers, and essays (Dikli, 2003). Due to the rapid improvement in technology, another type of testing has gained popularity over the last few years: computer-based testing. This type of testing promises greater efficiency, security, and immediacy of scoring (McFadden, Marsh, & Price, 2001). In traditional tests, the items are presented linear; one at the time, in the same order, to all students. Linear tests can be used in computer-based testing, but it also creates the opportunity for adaptive testing. In adaptive tests, the order, type of items, and total number of items presented are changed, based on the abilities of the individual students (McFadden et al., 2001). One of the benefits of computer-based testing is that a student can receive immediate feedback (Bennett, 1997; Dikli, 2003). Currently, multiple-choice questions are used most often in computer-based tests, since these items can be easily scored by the automatic scoring mechanism. The combination of the increasing use of computer-based testing and the popularity of games in education has resulted in the emergence of digital game-based assessment.

Game-based assessment

Game-based assessment (GBA) is "the application of principles of game design to measure performance when people are striving to perform at their best" (Heinzen, 2014, pp. 1). In other words, elements of games are implemented in (computer-based) tests. Traditional principles of assessment influenced by the insight of game designers seems to be a promising mash-up to make assessment more authentic (Dunning, Heath, & Suls, 2004). As mentioned earlier, examples of game elements used in game-based learning are points, badges and leader boards. These game elements are suitable in game-based learning or in consecutive formative tests. However, if students take a test only once, not all these game elements might be suitable as some elements are solely used to seduce players into playing more often. Game elements that might be suitable for summative assessment contexts are the use of colours, moving elements, and different types of game interactions (e.g. clicking, swiping, aiming). The elements 'freedom' (or nonlinearity) and rewards are also suitable in more extensive forms of game-based assessment.

According to Heinzen (2014) there are three important things to keep in mind when designing or using game-based assessment. First, the focus in game-based assessment should only be on

assessment, not on finding new ways to learn. Learning is often a welcome by-product of game-based assessment, but it is not the main goal. Secondly, it is important to keep in mind that game-based assessment is not identical to a video game; the use of games in education should not be the goal itself, but it should be used as a tool to increase engagement and motivation. Third, game-based assessment must be engaging and voluntary in order to capture peak performance.

Game-based assessment can be used in summative assessment, but it is mostly used for formative assessment, since it is most suitable for monitoring performance by assessing the information trails that learners naturally leave behind when playing a game, such as their strategy (Heinzen, 2014). The information gathered through game-based assessment generally consists of four types of observations: time to respond, accuracy of answers, points earned and number of attempts (Heinzen, 2014). This information generates a great opportunity for giving feedback.

In game-based assessment, two design principles are combined: the design of a game that provides engagement and enjoyment, and the design of assessment items that provide evidence for learning. Also, the design of the assessment should be valid. It is a challenge to balance these design considerations to maximize the effectiveness of assessment without losing the game-like characteristics such as fun and engagement (Kim & Shute, 2015). The use of games in assessment contexts is relatively new. This means that the effects and validity of game-based assessment are relatively unknown.

Validity and known effects of game-based assessment

The possible effects of gamification on students' test scores tells us something about the validity of the concerning assessment, namely: does the assessment measure the abilities of students correctly? First, it is important to focus on what validity is and which validity measurements are commonly used in educational measurement. The definition of validity provided in the new 2014 *Standards* (AERA, APA, & NCME, 2014, p.11) is as follows:

> *Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.* Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself.

There are two different approaches to support validity of the test results: the standard approach (AERA, APA, & NCME, 2014) that has been used for many years, and the more recently introduced argument-based approach (Kane, 1992; Kane, 2013). In the standard approach to validity, different types of validity can be distinguished. One of these types of validity is content validity: the extent to which the items of a test reflect the content that should be measured. A controversial sub-measurement of content validity is 'face validity'. Researchers generally use the term to express that the findings look and feel right on the surface (Royal, 2016). In educational assessment, this is an important validity measurement

to consider: if a test does not feel right, teachers or educationalists will not use it. However, face validity is not seen as a scientifically justified part of validity evidence, since it is not an objective measurement of validity. There are also more objective types of validity, for example construct validity: the extent to which a test measures the skills or abilities that should be measured. Another objective measurement of validity is criterion validity. This type of validity assesses whether a test reflects a certain set of abilities. This means that the scores of the test should be a good predictor of non-test behaviour or outcome criteria and that the test scores should correlate with other measures of the same construct (Zumbo & Chan, 2014).

Another approach to support validity is the argument-based approach. In this approach, test score interpretations and uses are valid when they are clearly stated and supported by appropriate evidence (Kane, 2013). This appropriate evidence is collected through practical arguments that are critically evaluated. The evaluation takes place in forms of clarity, completeness, coherence of the network, and its plausibility of its inferences and assumptions (Kane, 1992; Kane, 2013). In this research, the focus will be on the standard approach to support validity, and specifically content and construct validity. No statements can be made about criterion validity, since there is no data available to compare the data of this research with.

Known effects of game-based assessment

Even though the effects of gamification are clear, the effects of game-based assessment are still relatively unknown. One aspect that is known to influence several aspects of game-based assessment is linearity: the degree to which a players' freedom or control is restricted (Warren, 2009). Nonlinearity often relates to high levels of enjoyment, since the player has more freedom to choose their own path within the game (Chen, 2007). However, linearity does not necessarily mean that players experience less enjoyment than in nonlinear play. For example, Kim and Shute (2015) found no significant difference in enjoyment between linear and nonlinear gameplay. Also, linearity forces players to follow a predetermined path, which makes it more likely to ensure content validity (Almond, Kim, Velasquez, & Shute, 2014). On top of that, linearity in gameplay sequences might increase reliability, especially when the game is not adaptive.

Despite the known relationship between motivation, linearity, gender, and gamification, there are still a lot of unclarities regarding the effects of gamification on the response behaviour of students in assessment. For example, it is unclear to what extent the test scores are influenced by these game-based elements. It is essential to find out to what extent the test scores are affected, as this provides us important information about the validity of the assessment. Also, the relationship between gamification and other response behaviours, such as response time and concentration is unclear. This research is carried out to find out to what extent gamification influences the response behaviour of students in English vocabulary multiple-choice assessment.

Research questions and model

This research question that this research aims to answer is: to what extent does gamification affect the response behaviour of students in an English vocabulary multiple-choice assessment? The game interaction in three game conditions is different, with an increasing difficulty: clicking, swiping, and shooting. Response behaviour includes four aspects: test score, response time, motivation, and concentration. The effect on the response behaviour, and especially the test score, will tell us something about the validity of the assessment. The relationship between the response behaviours and the personal characteristics gender and game experience will also be examined.

Based on previous research, it is unclear to what extent gamification will influence the test score of students, and with that the validity of the test. However, the researcher expects that students will be slightly distracted by the game. Therefore, it is expected that game-elements will negatively influence the number of correct answers. Because of this expected distraction, a positive relationship is expected between game-based elements and response time; game-elements will probably create distraction and therefore students will take more time to answer the question. Next, it is expected that the test score and response time are affected more as the difficulty of the game interaction increases. The possible distraction is also expected to negatively affect the concentration in the game conditions. Based on previous research, it is expected that students will be more motivated in the game conditions. Previous research indicates that boys have an advantage in the early stages of game-based learning. Since this is a short assessment, it is expected that this effect will also be found in this research and therefore boys will score higher than girls in the game conditions. Last, it is expected that experienced gamers will score higher than non-experienced gamers in the game conditions.

Scientific and practical relevance

Recent research mostly focused on the positive effects of games in education, such as the increase of student motivation and engagement, and on how games can be used in assessment. However, there is lack of information about the validity of game-based assessment. This research on the response behaviours of students, and especially the test scores, in game-based assessment will therefore contribute to the field of educational research. It will also contribute to the field of education, since the results can help experts to make an informed, substantiated decision to use game-based assessment in their daily practice.

Research design and methods

Research design

This study uses a quantitative experimental research design with a randomized controlled trial. In this research, game-based elements, gender, and game experience are the independent variables and test score, response time, motivation and concentration are the dependent variables. The dependent variable game-based elements consists of three experimental conditions, with an increasing interaction: clicking, swiping and shooting. An overview of the aspects of the four conditions can be found in Table 1. The games used in the experimental conditions will be compared to a traditional computer-based test. By comparing the test scores of the different conditions, it can be determined if construct validity is affected. The decisions made during the design process of the game and the items will be used to support content validity.

Table 1

Aspects of the four conditions of the experimental research design

	Traditional	**Clicking**	**Swiping**	**Shooting**
Interaction	Clicking	Clicking	Swiping	Shooting
Layout	Black and white	Colourful (space-themed)	Colourful (space-themed)	Colourful (space-themed)
Game objects	No	Yes	Yes	Yes
Rewards	No	Yes	Yes	Yes
Feedback	Afterwards	Direct	Direct	Direct

Respondents

In total, 439 students were invited to participate in this study. Due to illness and some minor technical issues, the final sample includes 405 students (192 girls and 213 boys) from the last grade of primary school and the first grade of secondary school, aged 10-13 (M = 11.60, SD = 0.61). Dutch urban and sub-urban schools from the region of Overijssel and Gelderland (N = 8) were selected to participate. An overview of the participating schools can be found in Appendix 1.

For both English and digital literacy, national core goals are established by the National Institute for Curriculum Development (SLO[1]). Since the curriculum of all schools is created based on these national core goals, it is assumed that there are no large differences between different regions in the Netherlands regarding these topics. Therefore, the schools in these regions should be a good sample of the population.

[1] http://tule.slo.nl/Engels/F-KDEngels.html

To minimize the effects of school-specific variables, the students were randomly assigned to either the control condition or one of the three experimental conditions by using a login code. This randomization crossed schools and classes, so that in each class there were students assigned to each condition. This way of randomization ensures internal validity, as it minimizes the possibility that the found effects can be explained by an extraneous variable. The random distribution of the students over the four conditions has led to four equal groups, as shown in Table 2.

Table 2

Distribution of students in the different conditions (N)

	Traditional	**Clicking game**	**Swiping game**	**Shooting game**	**Total**
Girls	48	50	56	38	192
Boys	56	51	44	62	213
Total	**104**	**101**	**100**	**100**	**405**

Instrumentation

Game design

The first instrument that is used, is an assessment tool in which a traditional computer-based test and three game-based assessment conditions are created. Below, the design process of this assessment tool is described and supported by literature.

Educational games can be broadly categorized into three groups: multimedia approaches tightly linked to content presentation, repurposing pre-existing games for education, and specially designed games that seek a balance between fun and educational content (Moreno-Ger, Burgos, Martínez-Ortiz, Sierra, & Fernández-Manjón, 2008). In this research, a game will be specially designed. According to Prensky (2001), this should be the key to success. However, reaching a balance between fun and learning in gameplay is proven to be hard. Despite that, there are many success stories of game designs that were able to engage players who were not interested in the educational content (Moreno-Ger et al., 2008).

Apart from the different games and purposes of these games, there are general guidelines in educational game design. According to the guidelines of Moreno-Ger et al. (2008), an appropriate genre is chosen, or in other words: creating a game environment that will be suitable for the population. To meet the interest from students best, 15 Dutch students, aged 9-12 were involved in determining the game genre and lay-out. In groups of three, they created their own game themes. After they presented their ideas, the themes were scored by the students. The lay-out theme that eventually has been chosen, is ‘space’, which was positively assessed by both girls and boys, and therefore seems to be very suitable for this project.

According to the guidelines of Shi and Shih (2005), different game factors are considered in designing the game. These game factors, as described in Table 3, are assessed in the pilot test. Only the

game factor ‘sociality’ was left out, since interaction between students is not possible and desirable in individual assessment. In this game design, the major difference between the three game conditions is the type of interaction (e.g. clicking, swiping, and shooting). Due to the available resources, the fact that students only play the game once, and the fact that the game must be compared to a traditional linear test, no freedom of choice was added in the game. If students only play a game once, the experience they gain is little, and therefore the risk of reducing motivation is minimal (Westrom & Shaban, 1992). Also, restricting the freedom ensures content validity and reliability, which are important aspects in this research (Almond et al., 2014; Chen, 2007; Kim & Shute, 2015).

Table 3

Game factors

Game factor	**Description**
Game goals	The main concept of the game design, on which all factor designs should be based The designer should think about the type of experience they want to provide The players pursue these goals
Game mechanism	Includes methods used to achieve designer goals and to ensure smooth functioning of the game
Interaction	All interactions and conflicts that occur between the game and the players
Freedom	The amount of actions that players can perform in the game and how much freedom of choice they have within these (individual) actions
Game fantasy	The game environment and background, in which fantasy is an important aspect Fantasy does not imply unrealistic elements
Narrative	Describes what occurs in the virtual world, verbal and/or in media
Sensation	Multimedia presentation of the virtual world
Game value	The game should attract players to launch the game
Challenges	The effort that players should put into the game to achieve goals
Sociality	The interaction between people through the game
Mystery	Creating a game environment that involves player curiosity and exploration

Note. Based on the descriptions from Shi and Shih (2015).

Apart from different game factors, different game objects can also be distinguished. Mavridis and Tsiatsos (2017) describe four types of game objects in an educational game, based on their function: question objects, information objects, assistant objects, and dummy objects. These game objects are described in Table 4. In this game, the question object is an astronaut that is asking for the translation of a word through a text balloon. The correct translation can be selected by choosing one of the aliens on the planets, as showed in Figure 1. There are a few different assistant objects, for example there is

an option to close the game. Also, at the end of the assessment, there is a screen in which the results of the test are presented and a button for entering the questionnaire is displayed. The astronaut functions as both an assistant object and a question object in the swiping and in the shooting condition. In these conditions, the astronaut must be swiped or shot to the right answer and therefore the astronaut is used as a navigator (you must give an answer, to be able to proceed). In the clicking condition, the astronaut is solely a question object.

Figure 1. Question objects in the game conditions

Information objects are used in the beginning of the game. A short introduction text shows what the students need to do in the specific condition. Dummy objects that are used, are space-themed items, such as stars, planets and lights. Direct feedback is implemented in the game by using rewarding objects. If students answer a question correctly, a lot of small colourful stars will appear. If their answer is incorrect, the astronaut will shake. If a student answers 5 consecutive items correctly, they will receive a reward in the form of a star. These types of objects take advantage of the natural process of social comparison and enable self-assessment and self-evaluation (Suls et al., 2002). In this research, social comparison was only possible in conversations after the test is finished, and therefore the rewards focused mainly on enabling self-assessment and self-evaluation.

Table 4

Game objects

Object	**Description**
Question objects	Objects that contain questions
Information objects	A display or other type of object that shows information about the game, for example the remaining time or the amount of questions that are left
Assistant objects	Objects that help students to move and navigate inside the virtual environment
Dummy objects	Objects that are decorative, and have no function apart from creating an attractive environment

Note. Based on the description of a game-based assessment by Mavridis and Tsiatsos (2017)

Game conditions

As mentioned in the research design, three different game conditions were created. In the clicking game, game-based elements are added to the test, but they do not influence the acts of the students. In this condition the students only have to click on the correct answer which is the same interaction as in the control condition. The swiping game makes use of the interaction 'swiping'; students have to swipe the astronaut to the correct answer, which is on the left or on the right. This influences the acts of the students but does not differ a lot from the act of clicking. In the shooting game, students have to shoot the astronaut to the correct answer by aiming and releasing. In all experimental conditions, the game-element "rewards" will be applied. A full overview of the different conditions can be found in Appendix 2. The user-friendliness of the final prototype was also tested by two students, aged 11 and 12, using the think aloud protocol. Both students did not understand why their actions did not have any effect. During the introduction, the actions are blocked so that the students must watch the whole instruction before they can proceed. Since this was unclear to the students, it is decided to mention this during the oral introduction. Also, they found the instruction for the shooting game unclear; they could not find the cause for the unsuccessful shooting interaction. Therefore, this introduction was slightly changed before the data collection started; two rectangles were created to make clear that the students only must use these areas to aim, instead of a large triangle. The final introduction can be found in Figure 2.

Figure 2. Instruction in the shooting condition

Items

The vocabulary items are selected through Wozzol[2], a website with a collection of vocabulary lists from different English teaching methods. In the Netherlands, students are expected to have an English proficiency level of A1 at the end of primary school. Therefore, an A1-level assessment was made by an English assessment specialist. To ensure content validity, this item list was pilot tested by 15 students aged 9-12. During this pilot test, it became clear that most students did know many translations without any hesitation and the average score was high. To be able to do meaningful analyses on the English ability of students in different conditions and to ensure content validity, a certain level of difficulty is needed. Therefore, it was decided to add more difficult items (A2-level) to the test. The

[2] https://www.wozzol.nl/

vocabulary list that is used in this research can be found in Appendix 3 and consists of 20 A1-level items and 10 A2-level items.

Questionnaire and logbook

The second instrument that is used, is a computer-based questionnaire, presented in Qualtrics[3]. This questionnaire consists of personal and evaluative questions that assess the variables gender, game experience, motivation, and concentration and was filled out by the students after they finished the test. A button to open this questionnaire was presented automatically at the end of the test. The numerical code the students used to login to the assessment tool is transferred to this questionnaire automatically, so that the data of the game could be linked easily to the questionnaire. The evaluative questions about motivation and concentration were assessed using a 5-point Likert scale (1 = totally disagree, 5 = totally agree). The questionnaire can be found in Appendix 4.

Last, a logbook is filled out by the researcher during the experiment. In this logbook important information about the test situation was documented, such as technical issues or other difficulties that could affect the outcomes. The logbook can be found in Appendix 5.

Procedure

Before the data collection started, the Ethics Committee of the University of Twente was asked for approval. The granted permission can be found in Appendix 6. Also, one week prior to the experiment, a passive consent form was distributed among students' parents explaining the nature of the study and the possibility to retract. There was no objection to participation from the invited students nor their parents. The passive consent form can be found in Appendix 7.

At the school, the researcher orally introduced the experiment by explaining the goal, content and procedure of the experiment, supported by a visual presentation in PowerPoint. The test was executed on a tablet, which was provided by the researcher for all participating students. Also, two 4G-NetGear adapters where used to create the same, stable internet connection in all schools. When the researcher finished the instruction, the login codes were randomly distributed amongst the students. The students typed in this code to start the assessment. There was no time-limit to finish the test. The students took the test together with students of their own class, at the same time, in the same classroom. After they finished the game, they were automatically asked to fill out the online questionnaire they were able to open by clicking on a button at the end of the game. Students were asked to remain quiet until the last student was finished.

Data analysis

The distribution of the number of correct answers and response time in the different conditions were examined with descriptive statistics. Next, the relationship between the use of game elements and the test scores was tested through an analysis of variance (ANOVA-test). Descriptive and statistical

[3] https://www.qualtrics.com

analyses were provided for the questionnaire. The relationship between gender, game experience, and the test scores is examined using an ANOVA-test. The relationship between motivation, concentration and test score was tested by conducting a Pearson's r. All statistical tests were two-sided. A *P* value of < 0.05 is considered statistically significant.

The reliability of the test is approximated by the Kuder-Richardson-20 analysis, since this analysis fits dichotomous items best. The results of this analysis show that the reliability of the test, based on the traditional test, is sufficient in this context; KR (20) = .684.

Results

In this section, the results of the data analysis are presented. First, the test score is described by analysing the different conditions. Also, the relationship between test score and gender is examined. Next, the average response time is presented per condition. After that, the effect of game experience on both test score and response time are presented. Last, an overview of the assessed aspects in the questionnaire are discussed, such as motivation, concentration and valuation of the test.

Test scores

The mean score of the test was 25.50 correct items out of 30 items, which is relatively high. This indicates that the test is generally perceived as easy. As shown in Figure 3, the data is relatively normally distributed, but slightly skewed to the right.

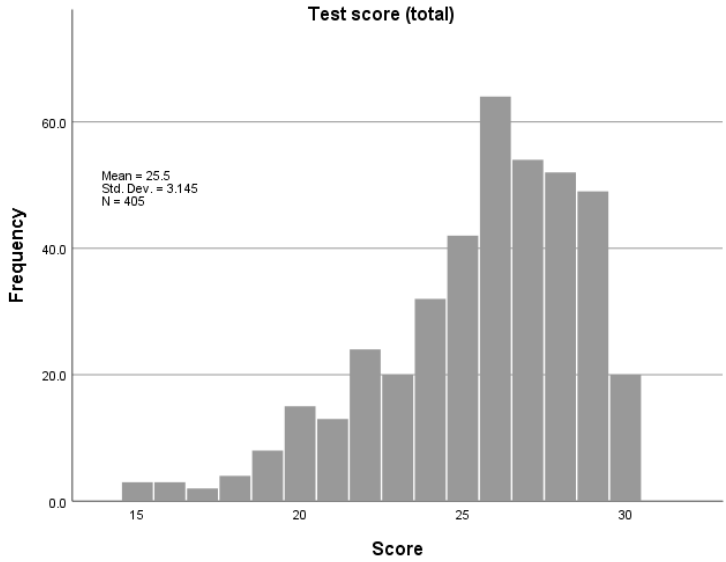


Figure 3. Distribution of the test scores

In Table 5, the test scores in the different conditions are displayed. The test scores are nearly equal. The shooting game shows the lowest mean score, while the swiping game shows the highest mean score. One-way ANOVA shows that the difference between the scores is considered statistically non-significant ($F(3,404) = .487$, $p = .691$).

Table 5

General overview of the test scores

Condition	N	M	SD
Traditional	104	25.59	2.77
Clicking game	101	25.53	3.32
Swiping game	100	25.70	3.08
Shooting game	100	25.19	3.41
Total		25.50	3.15

The results in Table 6 show that the overall mean score of boys is significantly higher than the overall mean score of girls ($F(1,404) = 5.502$, $p = .019$). As presented in this table, boys score higher in all four conditions; especially in the traditional test ($M_{difference} = 1.05$) and the swiping game ($M_{difference} = 1.26$). Although the overall difference is significant, the differences within the conditions between boys and girls were not significant.

Table 6

Difference between the test scores of boys and girls

Condition	**Boys**			**Girls**			**Mean difference**	**Significance**
	N	**M**	**SD**	**N**	**M**	**SD**		
Traditional	56	26.07	2.25	48	25.02	3.20	1.05	.053
Clicking game	51	26.16	2.79	50	24.90	3.71	1.26	.057
Swiping game	44	25.84	3.23	56	25.59	2.98	.25	.687
Shooting game	62	25.40	3.62	38	24.84	3.03	.56	.427
Total	192	25.85	3.02	213	25.12	3.24	.73*	.019

Note. The mean difference is significant at the .05 level

Response time

Regarding the response time, it must be considered that in the game conditions students have to wait to give an answer until the animations are finished. Therefore, the total test time of the game conditions is higher than the control condition. In every item, it takes 4.6 seconds to display the elements and give direct feedback. On top of that, the reward in the form of a star will take 4.4 seconds. If a student completes the test and has received all rewards, the total animation time is 2 minutes and 30 seconds. The average response time, as presented in Table 7, was measured from the moment that students were able to give an answer, to the moment where students gave the answer.

Table 7

Average response time per item per condition

Condition	**N**	**M**	**SD**
Traditional	104	3.43	1.67
Clicking game	101	3.18	.849
Swiping game	100	4.40	1.16
Shooting game	100	4.63	1.73

Note. Time in seconds

A one-way ANOVA shows that there is a significant difference between the response times in the different conditions ($F(3,404) = 25.886$, $p < .001$). Tukey post hoc test revealed that the response time was significantly higher in the swiping (4.40 ± 1.16 sec, $p < .001$) and the shooting game (4.63 ± 1.73 sec, $p = < .001$) compared to the traditional test (3.43 ± 1.67 sec). Also, Tukey post hoc revealed

that the response time was significantly higher in the swiping (4.40 ± 1.16 sec, $p < .001$) and the shooting game (4.63 ± 1.73 sec, $p = < .001$) compared to the clicking game (3.18 ± .849 sec). No significance difference has been found between the traditional test and the clicking game ($p = .596$) and between the swiping and shooting game ($p = .628$).

Game experience

In the questionnaire, the game experience of students was divided into five categories, as shown in Figure 4. The categories were based on an average of 15 hours per week, based on research reports in different online newspapers, such as the Daily Mail and the Volkskrant. (Belghmidi, 2019; Cahillane, 2018; Effting, 2016).

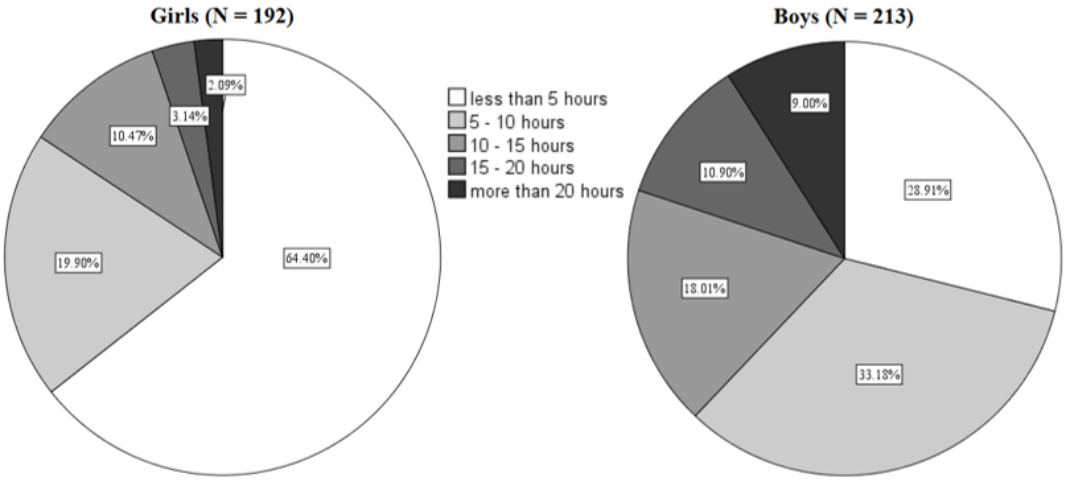


Figure 4. Average amount of gaming per week. Missing values: 3

During the data analysis it became clear that the chosen categories do not fit this group of participants; only 19.90% of the boys and 5.23% of the girls reported that their average game time is more than 15 hours per week. Therefore, it has been chosen to rearrange the categories into three new categories: unexperienced gamers (less than 5 hours), average gamers (5-10 hours), and experienced gamers (more than 10 hours). This resulted in a relatively equally distributed dataset, which makes it possible to execute more meaningful analyses. The distribution of students over the three new categories and the average scores of each group are reported in Table 8.

Table 8

Mean score based on game experience

Game experience		**Test score**	
	N	**M**	**SD**
Unexperienced	184	24.82	3.22
Average	108	25.97	2.93
Experienced	110	26.17	3.07

Note. Missing values: 3

The results in Table 8 show that the mean score of students increases, as their game experience increases. A one-way ANOVA shows that there is a significant difference between the test scores for students with different game experience ($F(2, 401) = 8.263$, $p < .001$). Tukey post hoc revealed that the test score was overall significantly higher for experienced gamers (26.17 ± 3.07, $p = .001$) and average gamers (25.97 ± 2.93, $p = .007$) than unexperienced gamers (24.82 ± 3.22). No significant difference was found between average and experienced gamers ($p = .882$). In Table 9, the average scores per condition are presented, in relation with game experience.

Game experience is of course only expected to influence the game conditions. However, a one-way ANOVA of the game conditions shows that the difference for game experience is less significant ($F(2,299) = 5.884$, $p = .003$). Tukey post hoc reveals that experienced gamers (26.19 ± 3.19, $p = .004$) score significantly higher than unexperienced gamers (24.75 ± 3.31) in the game conditions. No significant difference was found between average and experienced gamers ($p = .715$) and average and unexperienced gamers ($p = .057$).

When analysing the conditions individually, a one-way ANOVA for the clicking game shows that there is a significant difference between game experience and game score within this condition ($F(2,100) = 4.659$, $p = .012$). Tukey post hoc reveals that experienced gamers (26.45 ± 3.01, $p = .015$) score significantly higher than unexperienced gamers (24.27 ± 3.63). No significant difference was found between experienced gamers and average gamers ($p = .802$) and between average and unexperienced gamers ($p = .061$). No significant differences between game experience and game score were found within the swiping game ($F(2, 98) = 1.252$, $p = .291$) and within the shooting game ($F(2,99) = 1.542$, $p = .219$).

Remarkably, a one-way ANOVA for the traditional test shows that there is a significant difference between the game experience and game score within this condition ($F(2,101) = 3.150$, $p = .047$). However, Tukey post hoc reveals that there is no significant difference between unexperienced and experienced gamers ($p = .274$), unexperienced and average gamers ($p = .055$), and average and experienced gamers ($p = .890$) specifically.

Table 9

Mean score per condition for game experience

	Unexperienced			**Average**			**Experienced**		
	N	**M**	**SD**	**N**	**M**	**SD**	**N**	**M**	**SD**
Traditional	56	24.98	3.01	27	26.48	2.24	19	26.11	2.49
Clicking game	37	24.27	3.63	31	26.06	2.84	33	26.45	3.01
Swiping game	47	25.34	3.12	24	25.46	2.73	28	26.46	3.30
Shooting game	44	24.52	3.20	26	25.81	3.78	30	25.63	3.31

Note. Missing values: 3

The results showed some expected and unexpected effects of game experience on the test score. Also, game experience was found to affect the response time. A one-way ANOVA shows that there is an overall significant difference in response time between unexperienced, average and experienced gamers ($F(2, 401) = 3.661$, $p = .027$). Tukey post hoc reveals that the response time of experienced gamers (3.60 ± 1.43, $p = .019$) is significantly lower than the response time of unexperienced gamers (4.10 ± 1.52). No significant differences are found between average gamers and experienced gamers ($p = .302$) and between unexperienced and average gamers ($p = .554$).

Analysis within the conditions indicate that this difference is significant in the clicking game ($F(2,100) = 4.007$, $p = .003$). Tukey post hoc reveals that the response time of average gamers ($3.52 \pm .970$, $p = .002$) is significantly higher than the response time of experienced gamers ($2.81 \pm .553$). No other significant differences were found in this condition. In the shooting game, ANOVA-analysis shows that there is a significant difference in response time between unexperienced, average and experienced gamers ($F(2,99) = 10.686$, $p < .001$). Tukey post hoc for this condition reveals that the response time of average gamers (3.77 ± 1.21, $p < .001$) and experienced gamers ($4.20 \pm .288$, $p = .004$) is significantly lower than the response time of unexperienced gamers (5.44 ± 1.77). No significant differences were found between experienced and average gamers ($p = .566$). In the traditional test ($F(2,101) = 1.382$, $p = .256$) and the swiping game ($F(2,98) = .709$, $p = .495$), no significant differences were found. The average response times for game experience are displayed in Table 10.

Table 10

Average response time per condition for game experience

	Unexperienced			**Average**			**Experienced**		
	N	**M**	**SD**	**N**	**M**	**SD**	**N**	**M**	**SD**
Traditional	56	3.32	1.28	27	3.88	2.19	19	3.12	1.96
Clicking game	37	3.22	.850	31	3.52	.979	33	2.81	.553
Swiping game	47	4.43	.751	24	4.56	1.85	28	4.19	.997
Shooting game	44	5.44	1.77	26	3.76	1.21	30	4.20	1.58

Note. Time in seconds. Missing values: 3.

Questionnaire

As displayed in Table 11, students generally do appreciate the test. The traditional test has received the lowest grade and the variation in the grades is relatively high (M = 6.99, SD = 2.21). The average grade of the game conditions (M = 8.50, SD = 1.31) is a lot higher than the average grade of the traditional test. The swiping game is appreciated best (M = 8.58, SD = 1.13).

Table 11

Average appreciation of the test's look and feel

		Grade	
Condition	**N**	**M**	**SD**
Traditional	102	6.99	2.21
Clicking game	101	8.49	1.35
Swiping game	99	8.58	1.13
Shooting game	100	8.45	1.45

Note. Scale 1-10. Missing values: 3

As displayed in Figure 5, the traditional test is clearly most often (N = 21) rewarded with a grade below 6, which is indicated as insufficient. The game conditions are rarely valued as insufficient by the students. This indicates that students generally appreciated the game conditions more. Also, the traditional test has a high variance of given grades: all grades are given at least once. This indicates that the opinions of students about this test are very diverse. In the game conditions, the opinions of the students are less diverse and over 1/3 of the students rewards the test with a 9. No significant differences were found between the grades given by boys and girls ($F(1,401) = .049$, $p = .825$).

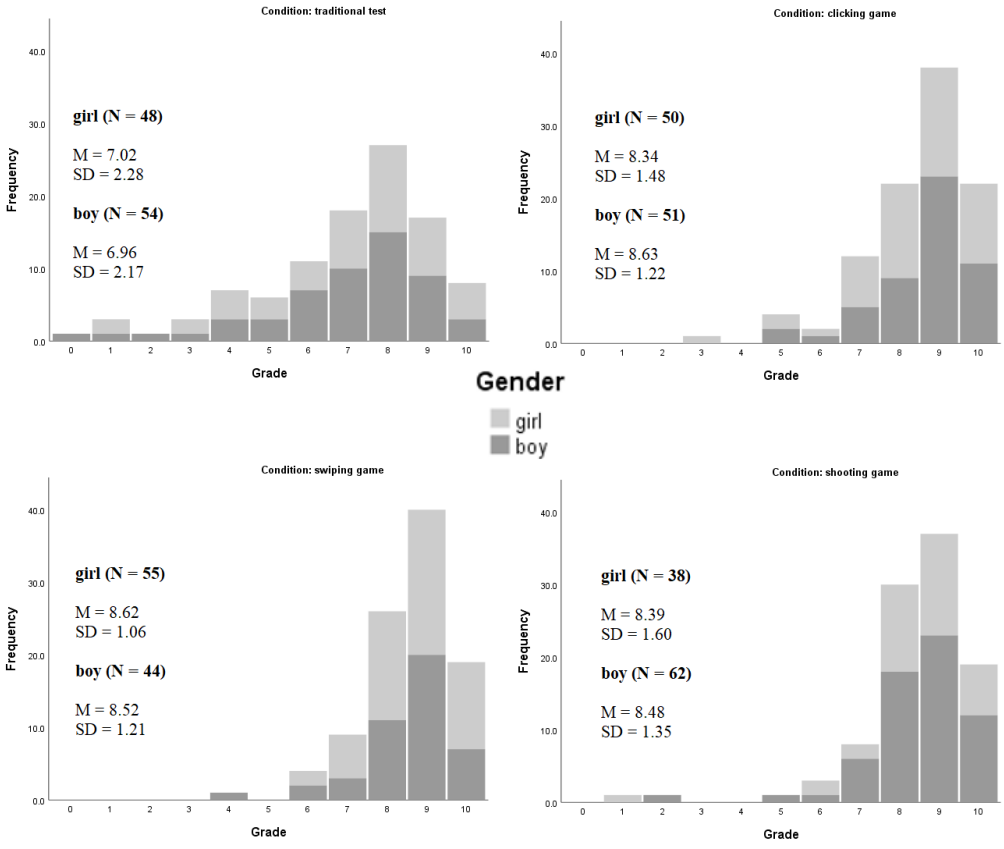


Figure 5. Distribution of given grades (appreciation) per condition. Missing values: 3

Motivation and concentration

In Table 12, the average motivation of the students is displayed based on self-evaluation of the statement "I was motivated during the test" on a 5-point Likert scale. The table shows that students were, on average, most motivated in the swiping game (M = 4.20, SD = .795) and least motivated in the traditional test (M = 3.32, SD = 1.17). A one-way ANOVA shows that students were significantly more motivated in the clicking game (4.08 ± .997, $p < .001$), swiping game (4.20 ± .795, $p < .001$) and shooting game (3.99 ± .927, $p < .001$) compared to the traditional test (3.32 ± 1.17). No significant differences have been found between the game conditions. As shown in Figure 6, there are no large differences between boys and girls regarding the reported motivation. An independent sample T-test indicates that these small differences were not significant ($F(1,401) = .328$, $p = .567$).

Table 12

Motivation: traditional condition versus game conditions

Condition	N	M	SD	Mean difference	Significance
Traditional	102	3.32	1.17		
Clicking game	100	4.08	.997	.756*	<.001
Swiping game	101	4.20	.795	.878*	<.001
Shooting game	99	3.99	.927	.666*	<.001
Total	402	3.90	1.04		

Note. Based on 5 points Likert scale (1 = totally disagree, 5 = totally agree). Mean differences are significant at the .05 level. Missing values: 3.

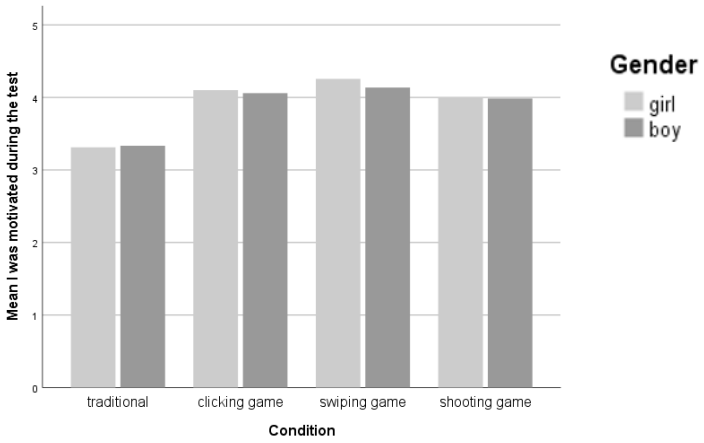


Figure 6. Motivation defined by gender. Missing values: 3

As displayed in Table 13, students also had the lowest average score for concentration in the traditional test (M = 4.08, SD = 1.01) based on self-evaluation of the statement "I was concentrated during the test" on a 5-points Likert scale. Students were, on average, most concentrated in the clicking game (M = 4.25, SD = .865). However, generally the students were highly concentrated in all conditions. A one-way ANOVA shows that there are no significant differences between the traditional test and the game conditions for concentration. No significant difference has been found between boys and girls regarding the reported concentration ($F(1,401) = .986$, $p = .321$).

Table 13

Concentration: traditional condition versus game conditions

Condition	**N**	**M**	**SD**	**Mean difference**	**Significance**
Traditional	102	4.08	1.01		
Clicking game	100	4.25	.865	.169	.569
Swiping game	101	4.22	.910	.144	.997
Shooting game	99	4.15	.936	.072	.880
Total	402	4.17	.931		

Note. Based on 5 points Likert scale (1 = totally disagree, 5 = totally agree). Mean differences are significant at the .05 level. Missing values: 3

The results show a positive correlation between motivation and test score, r = .171, n = 402, $p = .001$. As displayed in Table 16, this positive correlation is found in all game conditions. No significant relationship has been found for the traditional test.

Also, a Pearson's r was computed to assess the relationship between test score and concentration. The results of this analysis show a strong positive relationship between concentration and test score, r = .244, n = 402, $p < .001$. As shown in Table 14, this positive correlation is found in all conditions, except for the traditional test.

Table 14

Correlation with the test score for motivation and concentration in each condition

		Motivation		**Concentration**	
	N	**Pearson's r**	**Significance**	**Pearson's r**	**Significance**
Traditional	102	.090	.371	.099	.322
Clicking game	101	.208*	.037	.333*	.001
Swiping game	99	.213*	.035	.243*	.016
Shooting game	100	.247*	.013	.304*	.002

Note. Correlation is significant at the .05 level. Missing values: 3

Discussion, conclusion and recommendations

This research was set out to investigate to what extent gamification affects the response behaviour of students in an English vocabulary multiple-choice test, and with that the construct validity of the test. A successful within randomization was used to ensure internal validity. The results show no significant differences between the test scores of a traditional test and the test scores in the game conditions. This indicates that the interaction and design in the game conditions seem to have no influence on the measurement of the construct 'English vocabulary', and therefore construct validity is not affected. Even though some students reported that they had issues with the game interaction, especially in the shooting game, this is not reflected in the average test scores. These results are not in line with the hypothesis, since it was expected that the test score would be lower in the game conditions due to distraction and that these test scores would decrease as the difficulty of the game interaction increased. In the questionnaire, students indicated that they were, on average, highly concentrated in all conditions. This indicates that, based on these respondents, distraction does not play a role in game-based assessment.

In line with the hypothesis, a positive relationship was found between response time and the difficulty of the game interaction. The response time was significantly higher in the swiping and shooting games than in the traditional test and the clicking game. In the swiping and shooting games, the astronaut moves between two objects, which partly causes the higher response time; the speed of the astronaut is based on the speed of swiping and shooting. Therefore, it is unclear to what extent game elements affect the response time. Also, it must be considered that the total test time of the game conditions is higher, due to animations.

Next, it was expected that experienced gamers would score higher than unexperienced gamers in the game conditions. This hypothesis is partly supported. The results indicate that experienced gamers and average gamers generally score significantly higher than unexperienced gamers. When only analysing the game conditions, this difference is still significant, but only between experienced and unexperienced gamers. When looking in more detail, the significant differences are only found in the traditional test and the clicking game. This is remarkable, since these two conditions were expected to require no or minimal game experience. However, it must be considered that the compared groups are small in this analysis. Therefore, these results could be based on coincidence. Further research with larger group sizes is needed to explain this finding.

Also, the effect of game experience on response time was measured, which indicates a negative relationship between game experience and the response time in general. This means that the average response time decreases when the game experience increases. Further analysis showed that this negative relationship was found in the clicking game and the shooting game. No relationship was found in the traditional test and the swiping game. An explanation for this time difference might be that unexperienced gamers need more time to understand the game interaction or they might be more

distracted by the game-elements than experienced gamers.

Next, it was expected that boys would score higher in the game conditions than girls. This hypothesis is partly supported by the results; the test score of boys is higher than the test score of girls in all conditions, but only the overall difference was found to be significant. Previous research on game-based learning indicated that boys perform better than girls at the beginning of a game because they generally have more game experience, but this gap is closed to the point of statistically non-significant at the end of the game (Nietfeld et al., 2014). In this research, boys had on average more game experience than girls. Since a short test was used, it could be that the advantage boys might have in the beginning, due to previous game experience, still affects the score at the end.

Last, the results from the evaluative questions of the questionnaire showed that the game conditions were really appreciated by the students. The students did like the swiping game best, while the traditional test was clearly valued lowest. Also, students were more motivated in the game conditions than in the traditional test. No significant differences were found for concentration. A positive relationship has been found between motivation and test score and concentration and test score in the game conditions. No effects of motivation and concentration were found in the traditional test. These results indicate that motivation and concentration are important aspects in game-based assessment.

In conclusion, no evidence was found that gamification affects the test scores of students, and therefore gamification does not seem to affect the construct validity of this vocabulary test. The response time was affected by the difficulty of the game interaction and can probably be explained by the movements of objects. Even though a game-based test will take more time, students are more motivated and appreciate the way of testing more than a black and white computer-based test. Overall, experienced gamers score higher than unexperienced gamers. However, this difference is only found in the traditional test and the swiping game. Since this analysis is based on small groups, further research is needed to find out to what extent game experience affects the test score. Also, boys score higher than girls in all conditions, but this difference is only significant when looking at the overall score. The difference between boys and girls could be related to game experience, since boys have, on average, more game experience than girls. Generally, these results indicate that gamification does not influence construct validity of a test and therefore is a promising method to increase student motivation in testing. However, more research is needed to analyse unexpected results before this promising method will be used on a large scale.

Limitations and recommendations

The reliability of the research is impacted by the fact that the test was generally viewed as easy. Therefore, the test did not use a lot of children's cognitive capability, which may affect the results. Further research on more difficult cognitive tasks and gamification is needed, to indicate to what extent gamification affects the response behaviours of students, and with that the construct validity of a test. The generalizability of the results is limited by the fact that a computer-based test is used as a control condition. It was expected that students were used to this way of testing, however, during the data

collection it became clear that this was already a new approach to assessment for some children. Therefore, further research is needed to find out to what extent the response behaviour of students is affected by gamification in comparison with paper-based tests. Next, this research did not focus on criterion validity, since a completely new test was created, and this test was only conducted once. Therefore, no statements can be made about the extent to which game-based assessment affects criterion validity.

Also, the students took the test in their own classroom, with their school tables apart, to create a realistic test situation. In this setting, students are not able to read the questions on the screens of other students, but they were able to see what assessment condition their classmates were taking. Even though the researcher paid attention to the fact that students should focus on their own test, it could possibly have caused distraction for some students. To prevent this, further research could focus on individual test sessions or use other tools to prevent students from looking at the screens of their classmates. Also, this problem can be solved by dividing the conditions over whole classes. However, then this would negatively affect the internal validity.

Another limitation is that motivation and concentration were assessed using self-evaluation. To more convincingly assess the differences in motivation and concentration in different tests, it is advised to combine the self-evaluation with an objective observation of these variables. Last, further research is needed to find out to what extent game experience affects the test score in game-based testing, since the results of this research were unexpected.

References

Almond, R.G., Kim, Y.J., Velasquez, G., & Shute, V.J. (2014). Focus article: how task features impact evidence from assessments embedded in simulations and games. *Measurement: Interdisciplinary Research and Perspectives¸12*(1-2), 1-33. doi:10.1080/15366367.2014.910060

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association. Retrieved from https://eric.ed.gov/?id=ED565876

Armstrong, M.B., Ferrell, J., Collmus, A., & Landers, R. (2016). Correcting misconceptions about gamification of assessment: more than SJTs and badges. *Industrial and Organizational Psychology, 9*(3), 671-677. doi: 10.1017/iop.2016.69

Aydin, C. (2019). Test anxiety: gender differences in elementary school students. *European Journal of Educational Research, 8*(1), 21-30. doi: 10.12973/eu-jer.8.1.21

Belghmidi, L. (2019, June 17) 11 uur per week, zoveel gamen jongeren gemiddeld. *VRT*. Retrieved from https://www.vrt.be/vrtnws/nl/2019/06/14/11-uur-per-week-zoveel-gamen-jongeren-gemiddeld/

Bennett, R.E. (1997). *Reinventing assessment: speculations on the future of large-scale educational testing.* Princeton, NJ: ETS (Educational Testing Service). Retrieved from https://www.researchgate.net/publication/239064135

Bennett, R.E. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy, & Practice, 18*(1), 5-25. doi: 10.1080/0969594X.2010.513678

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31. doi:10.1007/s11092-008-9068-5

Buckley, P. & Doyle, E. (2016). Gamification and student motivation. *Interactive learning environments, 24*(6), 1162-1175. doi: 10.1080/10494820.2014.964263

Cahillane, M. (2018, June 15). Parents voice concerns about sexual imagery and violence in video games as study reveals children play 15 HOURS a week (yet almost half let kids access content above their age rating). *Daily Mail Online.* Retrieved from https://www.dailymail.co.uk/sciencetech/article-5850109/Children-play-average-15-hours-video-games-week-parents-concerned-violence.html

Chandel, P., Dutta, D., Tekta, P., Dutta, K., & Gupta, V. (2015). Digital game-based learning in computer science education. *CPUH Research Journal, 1*(2), 33-37. doi: 10.1016/j.compedu.2008.06.004

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM, 50*(4), 31-34. doi: 10.1145/1232743.1232769

Denden, M., Tlili, A., Essalmi, F., & Jemni, M. (2018). Implicit modelling of learners' personalities in a game-based learning environment using their gaming behaviors. *Smart Learning Environments, 5*(1), 1-19. doi: 10.1186/s40561-018-0078-6

Dikli, S. (2003). Assessment at a distance: traditional vs. alternative assessments. *The Turkish Online Journal of Educational Technology, 2*(3), 13-19. Retrieved from https://www.researchgate.net/publication/239585503_Assessment_at_a_distance_Traditional_vs_Alternative_Assessments

Dominguez, A., Saenz-de-Navarette, J., de-Marcos, L., Fernandez-Sanz, L., Pages, C., & Martinez-Herraiz, J. (2013). Gamifying learning experiences: practical implications and outcomes. *Computers and Education, 63*, 380-392. doi: 10.1016/j.compedu.2012.12.020

Dunning, D., Heath, C., & Suls, J.M. (2004). Flawed self-assessment. *Psychological Science in the Public Interest, 5*(3), 69-106. doi: 10.1111/j.1529-1006.2004.00018.x

Effting, M. (2016, April 5) Een op de tien jongens tussen 12 en 15 is gameverslaafd. *De Volkskrant.* Retrieved from https://www.volkskrant.nl/wetenschap/een-op-de-tien-jongens-tussen-12-en-15-is-gameverslaafd~b4e5d741/

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. doi: 10.3102/003465430298487

Heinzen, T.E. (2014, august 24). *Game-based assessment: two practical justifications*. Paper presented at KDD Annual Computer Science Conference, New York. doi: 10.13140/2.1.3251.7441

Inspectie van het Onderwijs (2019). *De Staat van het Onderwijs.* Retrieved from https://www.onderwijsinspectie.nl/onderwerpen/staat-van-het-onderwijs

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535. doi: 10.1177/1356389011410522

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448-457. Retrieved from https://www.researchgate.net/publication/281495676_The_Argument-Based_Approach_to_Validation

Kim, Y.L. & Shute, V.J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers and Education, 87*, 340-356. doi: 10.1016/j.compedu.2015.07.009

Lewis, R.S., Nikolova, A., Chang, D.J., & Weekes, N.Y. (2008). Examination stress and components of working memory. *Stress, 11*(2), 108-114. doi: 10.1080/10253890701535160

Lievens, F., & De Soete, B. (2012). Simulations. In S. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 383-410). New York, NY: Oxford University Press. Retrieved from https://www.researchgate.net/publication/294283938_Simulations_for_Personnel_Selection_An_Introduction

Mavridis, A. & Tsiatsos, T. (2017). Game-based assessment: investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning, 33*(2), 137-150. doi: 10.1111/jcal.12170

McFadden, A.C., Marsh, G.E., & Price, B.J. (2001). Computer testing in education. *Computers in the Schools, 18*(2-3), 43-60. doi: 10.1300/J025v18n02_04

McGonigal, J. (2011). *Reality is broken: why games make us better and how they can change the world.* New York, NY: The Penguin Press. Retrieved from https://hci.stanford.edu/courses/cs047n/readings/Reality_is_Broken.pdf

Moreno-Ger, P., Burgos, D., Martínez-Ortiz, I., Sierra, J.L., & Fernández-Manjón, B. (2008). Educational game design for online education. *Computers in Human Behavior, 24*(6), 25302540. doi: 10.1016/j.chb.2008.03.012

Nietfeld, J.L., Shores, L.R., & Hoffmann, K.F. (2014). Self-regulation and gender within a game-based learning environment. *Journal of Educational Psychology, 106*(4), 961-973. doi: 10.1037/a0037116

Panayiotis, A. & James, M. (2014). Exploring formative assessment in primary school classrooms: developing a framework of actions and strategies. *Educational Assessment, Evaluation an Accountability, 26*(2), 153-176. doi: 10.1007/s11092-013-9188-4

Papastergiou, M. (2009). Exploring the potential of computer and video games for health and physical education: a literature review. *Computers and Education, 53*, 603-622. doi: 10.1016/j.compedu.2009.04.001

Park, J., Chung, S., An, H., Park, S., Lee, C., Kim, S.Y., … Kim, K.S. (2012). A structural model of stress, motivation, and academic performance in medical students. *Psychiatry Investigation, 9*(2), 143-149. doi: 10.4306/pi.2012.9.2.143

Prensky, M. (2001). *Digital game-based learning*. New York: McGraw-Hill. doi: 10.1145/950566.950567

Prince, J.D. (2013). Gamification. *Journal of Electronic Resources in Medical Libraries, 10*(3), 162-169. doi: 10.1080/15424065.2013.820539

Royal, K. (2016). "Face validity" is not a legitimate type of validity evidence! *The American Journal of Surgery, 212*(5), 1026-1027. doi: 10.1016/j.amjsurg.2016.02.018

Shi, Y. & Shih, J. (2015). Game factors and game-based learning design model. *International Journal of Computer Games Technology, 2015*(1), 1-11. doi: 10.1155/2015/549684

Suls, J., Martin, R., & Wheeler, L. (2002). Social comparison: why, with whom, and with what effect? *Current Directions in Psychological Science, 11*(5), 159-163. doi: 10.1111/1467-8721.00191

Ventura, M. & Shute, V.J. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior, 29*(6), 2568-2572. doi: 10.1016/j.chb.2013.06.033

Vogel, J.J., Vogel, D.S., Cannon-Bowers, J., Bowers, C.A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: a meta-analysis. *Journal of Educational Computing Research, 34*(3), 229-243. doi: 10.2190/FLHV-K4WA-WPVQ-H0YM

Warren, L. (2009, September 18). *What do we mean when we say non-linear?* [Web log post]. Retrieved from http://digitalkicks.wordpress.com/2009/09/18/what-do-we-mean-when-we-say-non-linear/

Westrom, M. & Shaban, A. (1992). Intrinsic motivation in microcomputer games. *Journal of Research on Computing in Education, 24*(4), 433-445. doi: 10.1080/08886504.1992.10782018

Zumbo, B.D. & Chan, E.K.H. (2014). *Validity and validation in social, behavioral, and health sciences*. (Social indicators research series, volume 54). Cham, Switzerland: Springer. doi: 10.1007/9783-319-07794-9

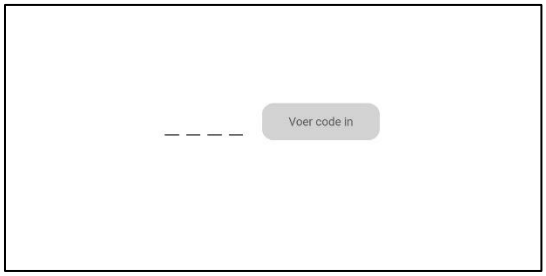
Appendices

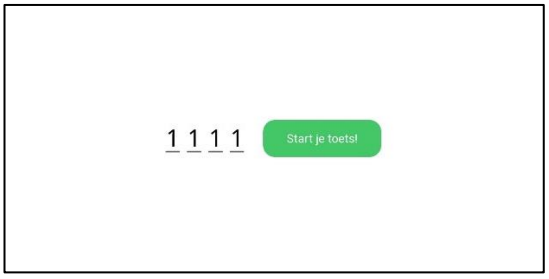
Appendix 1: Participating schools

School	Education	City	Number of students (invited)
Deventer Leerschool	Public primary school	Deventer (OV)	27
Het Erasmus	Public secondary school	Almelo (OV)	232
IBS de Tulp	Islamic primary school	Hengelo (OV)	12
Montessori van Lith	Montessori primary school	Deventer (OV)	30
OBS Beekbergen	Public primary school	Beekbergen (GL)	19
OBS Berg en Bos	Public primary school	Apeldoorn (GL)	25
OBS De Weier	Public primary school	Almelo (OV)	59
OBS Kolmenscate	Public primary school	Deventer (OV)	35

Appendix 2: Full overview of the assessment conditions

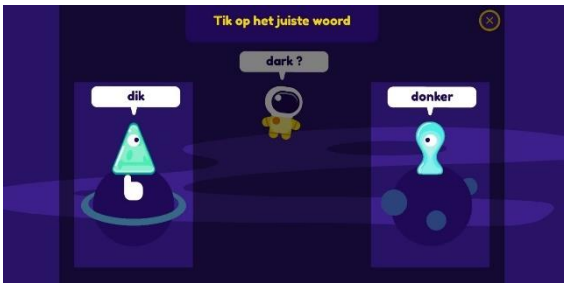
1. *Login*





2. *Traditional test*

3. *Clicking game*

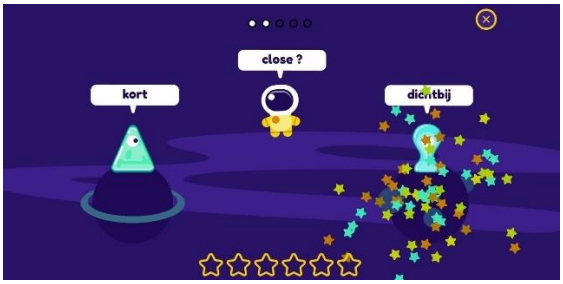


4. *Swiping game*

5. *Shooting game*

6. *Other*

Reward (direct feedback)



Reward (5 correct items in a row)

Quit game

Appendix 3: English vocabulary list

English	Dutch translation	Incorrect answer	Level
catch	vangen	botsen	A1
close	dichtbij	kort	A1
mountain	berg	land	A1
quick	snel	moeilijk	A1
question	vraag	oefening	A1
joke	grap	pijn	A1
ask	vragen	zeggen	A1
tired	moe	zwaar	A1
famous	beroemd	rijk	A1
body	lichaam	vriend	A1
early	vroeg	laat	A1
clever	slim	kleverig	A1
window	raam	regen	A1
look	kijken	bakken	A1
stay	blijven	wassen	A1
garden	tuin	eiland	A1
drive	rijden	fietsen	A1
journey	reis	dag	A1
owner	eigenaar	alleen	A1
waiter	ober	geduld	A1
meaning	betekenis	mening	A2
cottage	zomerhuis	katoen	A2
handy	handig	handen	A2
weigh	wegen	vliegen	A2
scream	gillen	rennen	A2
cinema	bioscoop	zwembad	A2
sailing	zeilen	surfen	A2
push	duwen	trekken	A2
fog	mist	kikker	A2
thunder	donder	tunnel	A2

Appendix 4: Questionnaire

Gamification

Bedankt voor het meedoen aan het onderzoek. Ik heb nog een paar vragen voor je. Alle antwoorden die je geeft, zullen anoniem verwerkt worden.

Persoonlijke informatie

Hieronder volgen een aantal persoonlijke vragen

Vraag 1 Wat is je leeftijd?

- ○ 9
- ○ 10
- ○ 11
- ○ 12
- ○ 13
- ○ 14

Vraag 2 Wat is je geslacht?

- ○ meisje
- ○ jongen

Vraag 3 Welk onderwijs volg je?

- ○ basisschool
- ○ vmbo
- ○ mavo/havo
- ○ havo/vwo
- ○ gymnasium

Vraag 4 Hoeveel uur per week speel je games?
Dit kan zijn op een console (Playstation, Xbox etc.), maar ook via een app (smartphone, tablet etc.)

- ○ minder dan 5 uur
- ○ 5 - 10 uur
- ○ 10 - 15 uur
- ○ 15 - 20 uur
- ○ meer dan 20 uur

Toetservaring
Hieronder volgen een aantal vragen over hoe je de toets ervaren hebt

---

Vraag 5 Was het duidelijk hoe je de toets moest maken?

- ○ Ja
- ○ Nee, want ... _____

Vraag 6 Geef aan of je het eens of oneens bent met onderstaande stellingen

	helemaal oneens	oneens	neutraal	eens	helemaal eens
Ik was gemotiveerd tijdens het maken van de toets	○	○	○	○	○
Ik was geconcentreerd tijdens het maken van de toets	○	○	○	○	○

Vraag 7 Welk cijfer geef je de toetsversie die jij gemaakt hebt?

0 1 2 3 4 5 6 7 8 9 10

Vraag 8 Wil je verder nog iets kwijt?

_____

Extra vraag 3b Welk niveau verwacht je te gaan doen op de middelbare school?

- ○ praktijkonderwijs
- ○ vmbo
- ○ mavo/havo
- ○ havo
- ○ havo/vwo
- ○ vwo
- ○ gymnasium
- ○ weet ik (nog) niet

Appendix 5: Logbook

Date	School	Remarks
31-10-2019	OBS Kolmenscate	- Login did not work for three students; these students were given a new login code. - Five students experienced crashing or freezing of the game. Three of them were not able to complete the assessment due to this issue. Additional remark: the technical issues seem to have something to do with the number of students participating at the same time (35). These technical issues were not experienced at such a large scale later.
31-10-2019	OBS Beekbergen	- Two students with the shooting game mentioned that the astronaut went the wrong way multiple times.
01-11-2019	Montessori van Lith	- Two students accidently closed the game without filling in the questionnaire. They filled in the questionnaire manually through a hyperlink. Their login code and the time they finished the questionnaire were written down, so that the data can be linked. - One student accidentally filled in the questionnaire twice (once partly). This code is written down, so that the incomplete response can be deleted later.
06-11-2019 07-11-2019 08-11-2019	Het Erasmus	- One student mentioned that a correct answer was marked as wrong during the gameplay. - A few students experienced the astronaut going the wrong way in the active condition when they started swiping too early. - Multiple students found the shooting interaction in the shooting game difficult. Some of them did not understand how to aim and shoot, while others experienced the astronaut going in the wrong way.
13-11-2019	OBS Berg en Bos	- Two students accidentally closed the game because their “swipe-movement” was too long.
14-11-2019	OBS de Weier	- One student (traditional test) mentioned that she accidentally made some mistakes because the new item immediately shows up after the answer is given. If she accidentally clicked twice, the answer was already given and unchangeable for the next question.
18-11-2019	IBS de Tulp	- One student accidentally closed the game but was able to login again. - Two students found the level of English too difficult.
18-11-2019	Deventer Leerschool	- It was not completely quiet during the test, due to the enthusiasm of a few students. - Three students accidentally closed the game in the swiping condition. Two of them were able to proceed, while one of them did not finish the game. - Two students experienced the astronaut going the wrong way in the shooting game

Appendix 6: Approval of the Ethics Committee

APPROVED BMS EC RESEARCH PROJECT REQUEST

Dear researcher,

This is a notification from the BMS Ethics Committee concerning the web application form for the ethical review of research projects.

Requestnr. :	191003
Title :	Validity of gamification in assessment
Date of application :	2019-06-25
Researcher :	Grobben, M.M.
Supervisor :	Veldkamp, B.P.
Commission :	Lubbe, R.H.J. van der
Usage of SONA :	N

Your research has been approved by the Ethics Committee.

The ethical committee has assessed the ethical aspects of your research project. On the basis of the information you provided, the committee does not have any ethical concerns regarding this research project.

It is your responsibility to ensure that the research is carried out in line with the information provided in the application you submitted for ethical review. If you make changes to the proposal that affect the approach to research on humans, you must resubmit the changed project or grant agreement to the ethical committee with these changes highlighted.

Moreover, novel ethical issues may emerge while carrying out your research. It is important that you re-consider and discuss the ethical aspects and implications of your research regularly, and that you proceed as a responsible scientist.

Finally, your research is subject to regulations such as the EU General Data Protection Regulation (GDPR), the Code of Conduct for the use of personal data in Scientific Research by VSNU (the Association of Universities in the Netherlands), further codes of conduct that are applicable in your field, and the obligation to report a security incident (data breach or otherwise) at the UT.

-

Appendix 7: Passive consent form

Beste ouder(s)/verzorger(s),

Mijn naam is Maaike Grobben. Ik volg de master Educational Science and Technology aan de Universiteit Twente. Voor mijn masterscriptie doe ik onderzoek naar de betrouwbaarheid van het inzetten van games in toetsen. Hiervoor zal ik op __*datum*__ eenmalig een meerkeuzetoets Engels afnemen in de klas van uw zoon/dochter. Naast het maken van de toets op een tablet, wordt er in een korte vragenlijst gevraagd naar het geslacht, de leeftijd, het onderwijsniveau en de game-ervaring van de leerling. Ook wordt er in de vragenlijst gevraagd hoe de leerling de toets heeft ervaren. Alle data zal **anoniem** verwerkt worden.

Via deze weg vraag ik u en uw zoon/dochter om toestemming voor:

- het verzamelen en verwerken van zijn/haar gegevens;
- het archiveren van de data;
- het geanonimiseerd publiceren van de data

Indien u of uw zoon/dochter **geen** toestemming geeft voor het deelnemen aan bovenstaand onderzoek, kan u dit doorgeven via het volgende mailadres: m.m.grobben@student.utwente.nl. Graag hierin de naam, klas en school van de betreffende leerling benoemen.

Ook voor andere vragen/opmerkingen met betrekking tot het onderzoek kunt u mij bereiken via bovenstaand mailadres.

NB. Deelnemers mogen te allen tijde de gegeven toestemming intrekken en zijn op ieder moment vrij om te stoppen met het onderzoek zonder daarvoor een reden te geven.