Vakgroep Communicatiewetenschappen

Faculteit politieke en sociale wetenschappen

Academiejaar 2016-2017

Digital Game-Based Learning Under The Microscope.

Development of a procedure for assessing the effectiveness of educational games aimed at cognitive learning outcomes

Anissa All

Promotor: Prof. Dr. Jan Van Looy

Co-promotor: Dr. Elena Patricia Nuñez Castellar

Imec – MICT – Ghent University

Proefschrift ingediend tot het behalen van de graad Doctor in de Communicatiewetenschappen

All, A. (2016). Digital Game-Based Learning Under The Microscope. Development of a procedure for assessing the effectiveness of educational games aimed at cognitive learning outcomes (Doctoral dissertation). Faculty of political and social sciences, Ghent University, Belgium.

Print: University Press

Cover: Nane Van Damme

Dit proefschrift werd gefinancierd door het Vlaams Agentschap voor Innovatie door Wetenschap en technologie (IWT)

Samenstelling Examencommissie:

Promotoren:	Prof. Dr. Jan Van Looy Vakgroep Communicatiewetenschappen Universiteit Gent
	Dr. Elena Patricia Nuñez Castellar Vakgroep Data-analyse Universiteit Gent
Voorzitter	Prof. Dr. Hans Verstraeten Vakgroep Communicatiewetenschappen Universiteit Gent
Leden:	Prof. Dr. Arnaud Szmalec Faculteit psychologische en pedagogische wetenschappen Université Catholique de Louvain
	Prof. Dr. Igor Mayer Academy of Digital Entertainment NHTV Breda University of Applied Sciences
	Prof. Dr. Lieven De Marez Vakgroep Communicatiewetenschappen Universitei Gent
	Prof. Dr. Tammy Schellens Vakgroep pedagogie Universiteit Gent
	Wim Govaerts CEO, Epyc for e-learning

Dankwoord

Na vier jaar is het eindelijk zo ver: dat doctoraat ligt er. De 18-jarige ik die vol enthousiasme aan de opleiding Communicatiewetenschappen begon, had nooit gedacht dat ik hier vandaag zou staan. Ik begon dan ook aan de opleiding met het doel journalistiek of communicatiemanagement te doen. Echter, in het derde jaar kreeg ik les van ene professor De Marez, meer bepaald Nieuwe Communicatietechnologieën. Wanneer ik zijn afstudeerrichting 'Nieuwe Media & ICT' bekeek waarin ik het vak 'Innovatieonderzoek' kon volgen, was ik plotseling gewonnen voor onderzoek naar Nieuwe Media. Bij deze zou ik dus graag eerst een woord van dank uiten aan Prof. Dr. Lieven De Marez voor deze lessen die hij vol enthousiasme en passie gaf en een opleiding aan te bieden die een zeer mooie brug bouwt tussen academisch onderzoek en de bedrijfswereld.

Toen ik mijn stage begon bij MICT -ik was echt zo enthousiast voor innovatieonderzoek dat ik dus ook graag eens de real deal wou ervaren- werd ik in ineens in de wereld van het serious game-onderzoek gegooid. Na mijn stage, kon ik ook aan de slag bij MICT voor een serious game project. Bij deze dus ook een speciaal woord van dank aan mijn promotor Jan Van Looy om me warm te maken voor dit type onderzoek en mij er te van overtuigen toch eens voor een doctoraatsbeurs te proberen gaan. Ook wil ik Jan, samen met mijn copromotor Dr. Elena Patricia Nuñez Castellar bedanken om -soms tot vervelens toe- mij tot het uiterste te drijven en me 200% te geven om toch mee te kunnen gaan naar ICA of die A1 publicatie toch te halen. Jullie hebben me geleerd om in mezelf te geloven, maar ook om kritisch te zijn ten opzichte van mezelf.

Verder wil ik graag al mijn collega's bij MICT bedanken voor de ongelooflijk leuke werkomgeving waar ik 5 jaar heb in mogen vertoeven, maar ook om dergelijke collega's te zijn die probleemloos online formulieren invullen wanneer ik weer eens stemmen aan het ronselen was voor één of andere wedstrijd. Lotte, Jasmien, Evelien D'heer, Matthias, Elena, Hadewijch & Evelien De Waele De Guchtenaere wil ik ook nog eens speciaal bedanken voor de leuke intermezzo's op kantoor, maar ook om een luisterend oor te bieden als er doctoraat gerelateerde frustraties de kop op kwamen steken.

Paulien, Steven, Ellen, Elke & Tom, bedankt voor de tijd aan de Universiteit zo ongelofelijk leuk te maken. Verder wil ik al mijn familie en vrienden bedanken voor de steun die jullie me de afgelopen jaren hebben geboden. Hier ook een speciaal woord van dank aan Dagmar, Inge & Nicky om mij met rust te laten wanneer ik besloot mij op te sluiten om dat doctoraat af te werken, maar ook om er als eerste terug te staan wanneer ik er me toch niet aan kon houden. Fiona, bedankt om mij -letterlijk- elke dag van mijn doctoraat bij te staan, mij de laatste maanden te steunen wanneer ik begon te twijfelen aan mezelf en buiten de werkomgeving er ook altijd te zijn voor mij. Verder wil ik mijn ouders bedanken om altijd in me te geloven en mogelijks nog enthousiaster te zijn dan mij wanneer ik aanvaard werd voor een conferentie of wetenschappelijk tijdschrift. Zonder jullie had ik hier vandaag niet gestaan, want jullie hebben mij geleerd dat ik alles kan doen wat ik wil, zolang ik er maar mijn best voor doe. En last but not least, Jef, mijn beste vriend, mijn lief en binnenkort mijn man, bedankt om er dag in dag uit te zijn voor mij en de persoon te zijn waarbij ik kan thuiskomen en alle doctoraatsstress in het niks verdwijnt. Wat er ook na dit doctoraat uit de bus valt, met jou naast mijn zijde, weet ik dat ik sowieso de gelukkigste vrouw ter wereld zal zijn.

Table of contents

Executive Summary: Nederlands

1. Introductie

Heterogeniteit in effectiviteitsonderzoek naar educatieve videogames en vragen die gesteld worden met betrekking tot de validiteit en betrouwbaarheid van enkele van deze studie design kenmerken hebben geleid tot een nood naar een meer gestandaardiseerde aanpak. Dit doctoraat was gericht op de ontwikkeling van een gestandaardiseerde procedure om effectiviteit van educatieve videogames te meten die zich richten op cognitieve leeruitkomsten (i.e., gericht op kennisoverdracht en niet zozeer op de ontwikkeling van vaardigheden of attitude-/gedragsverandering). In de eerste fase werd een eerste versie van deze procedure ontwikkeld. Hiervoor werden de verschillende studie design kenmerken van reeds gepubliceerde (quasi-) experimentele studies die als doel hadden de effectiviteit van educatieve videogames gericht op cognitieve leeruitkomsten te onderzoeken in een eerste stap in kaart gebracht door middel van een systematische literatuurstudie (hoofdstuk 2). Daarna werd effectiviteit van educatieve videogames aan de hand van een requirements analyse met relevante stakeholdergroepen geconceptualiseerd en geoperationaliseerd (hoofdstuk 3). Vervolgens hebben we enkele richtlijnen gedefinieerd voor effectiviteitsonderzoek naar educatieve videogames door middel van expert interviews (hoofdstuk 4). In een tweede fase hebben we de haalbaarheid van de procedure onderzocht door middel van drie experimentele studies in de voornaamste sectoren waar educatieve videogames geïmplementeerd worden, waarbij de procedure als handleiding werd gebruikt. In een derde fase werd op basis van de haalbaarheidsstudies een tweede versie van de procedure ontwikkeld.

2. Samenvatting van de voornaamste bevindingen

2.1. *Issues met betrekking tot gepubliceerde effectiviteitsstudies*

De systematische literatuurstudie heeft drie voorname issues naar voor gebracht. Eerst en vooral werd aangetoond dat er een grote heterogeniteit bestaat tussen studies. De voornaamste oorzaken van deze heterogeniteit kunnen de uitkomsten significant beïnvloeden. Zo worden er verschillende activiteiten geïmplementeerd in de controlegroepen (e.g., geen of een andere educatieve interventie), worden er verschillende maten gebruikt om effectiviteit te meten (e.g., gepercipieerd leren, score op een test ontwikkeld voor onderzoek, examenscores, attitude, motivatie, etc.) en gebruikt men verschillende statistische technieken om

leeruitkomsten te kwantificeren (e.g., vergelijking van vooruitgang, van post-test scores, percentage vooruitgang bekijken, etc.). Een tweede issue zijn suboptimale studie designs als gevolg van 'confounds' of verstorende factoren in een onderzoeksdesign die kunnen leiden tot vertekende resultaten. De voornaamste confounds in effectiviteitsonderzoek naar educatieve videogames zijn a) het toevoegen van extra elementen aan de interventie (e.g., verplichte leesopdracht, debriefing sessie), b) de rol van (e.g., begeleiding, procedurele hulp, enkel aanwezig om toezicht te houden) en het type (vertrouwde leerkracht of onderzoeker) instructeur en c) het mogelijks optreden van een oefeneffect wanneer dezelfde test pre- en post- interventie wordt toegediend. Al deze verstorende factoren maken het moeilijk om te weten of positieve effecten het resultaat zijn van de game als medium. Een derde issue heeft betrekking tot op replicatiemogelijkheden door onvoldoende informatie die wordt meegegeven in gepubliceerde studies. Zo wordt weinig informatie meegegeven rond hoe de interventies geïmplementeerd werden (e.g., wie was er aanwezig gedurende de interventie? In welke context vond deze plaats? Was gameplay individueel? Etc.), hoe rekrutering gebeurd is, op basis van welke karakteristieken de condities gelijk gehouden werden, welke testen er gebruikt werden en indien er testen werden gebruikt die ontwikkeld werden door de onderzoekers zelf, hoe deze ontwikkeld werden.

2.2. *Effectiviteit van educatieve videogames: conceptualisatie en operationalisering*

De requirements analyse wijst erop dat effectiviteit van educatieve videogames een multidimensionaal construct is dat bestaat uit drie categorieën van gewenste uitkomsten: leeruitkomsten, motivatie-uitkomsten en efficiëntie-uitkomsten. Voor elke categorie van uitkomsten kunnen verschillende indicatoren gebruikt worden. Voor leeruitkomsten kan dit a) een verhoogde interesse in de leerinhoud, b) prestatie m.b.t. de leerinhoud (vb. op een test) en/of c) transfer van de leerinhoud naar de 'echte' wereld zijn. Indicatoren voor motivatie-uitkomsten zijn: a) het creëren van een leukere leerervaring in vergelijking met traditionelere instructieve media en/of b) de student gemotiveerder maken om een bepaalde leerinhoud aan te leren. Efficiëntie-uitkomsten kunnen verwijzen naar a) een reductie in tijd om een bepaalde materie aan te leren en/of b) een kostenefficiëntere manier bieden om een bepaalde materie aan te leren. Motivatie- en efficiëntie- uitkomsten zijn geen alleenstaande reden om te investeren in de ontwikkeling en implementatie van educatieve videogames. Hogere motivatie en efficiëntie- uitkomsten moeten steeds gerelateerd zijn aan gelijkaardige leeruitkomsten die met de huidige leermedia bereikt worden.

2.3. *Best practices om effectiviteit van educatieve videogames te meten*

In hoofdstuk 4 hebben we enkele potentiële verbeterpunten gedefinieerd in effectiviteitsonderzoek naar educatieve videogames. Deze zaken hebben betrekking op de implementatie van de interventie en onderzoeksmethodes gebruikt om effectiviteit te meten. Met betrekking tot implementatie van interventies in zowel de experimentele als controlegroep, werden enkele praktijken afgelijnd die liefst worden vermeden om verstorende effecten te reduceren (e.g., begeleiding van de instructeur, extra elementen die toegevoegd worden aan de interventie die inhoudelijke informatie bevatten). Ook werden praktijken afgelijnd die wel toegelaten zouden zijn, zoals procedurele hulp en een training sessie. Daarenboven werden enkele variabelen naar voren geschoven die relevant zijn om gelijk te houden tussen de experimentele en controleconditie, zoals instructietijd, instructeur, tijdsslot en dag. Met betrekking tot de methodes die gebruikt worden, zijn voorgestelde verbeterpunten gerelateerd aan het toewijzen van deelnemers aan condities (e.g., variabelen die overwogen dienen te worden bij matchen van deelnemers in groepen), algemeen design (vb. toevoegen van een follow-up studie) en testontwikkeling (vb. ontwikkeling en piloten van parallelle tests). Hoewel eerdere al suggesties werden gemaakt tot verbetering van effectiviteitsonderzoek naar educatieve videogames, omvatten deze niet elk aspect van het studie design, zoals welke variabelen het liefst vergelijkbaar zijn in de condities, de rol van de instructeur, elementen die liefst niet worden toegevoegd aan de implementatie van de interventie. Dit hoofdstuk heeft getracht om elk aspect van het studie design te behandelen.

2.4. *Empirische bevindingen met betrekking tot effectiviteit van educatieve videogames*

Buiten het testen van de haalbaarheid en optimalisatie van de procedure die ontwikkeld werd in de eerste fase, hebben deze effectiviteitsstudies ook enkele interessante bevindingen naar voor gebracht met betrekking tot effectiviteit van educatieve videogames. De eerste haalbaarheidsstudie (hoofdstuk 5) heeft aangetoond dat hoewel er geen significante verschillen gevonden werden tussen een game en een klassikale les om Engelse woordenschat te leren, de woordenschat langer bleef hangen bij leerlingen die de klassikale les kregen. De traditionele les bleek dus effectiever op langere termijn (i.e., na 3 weken). Dit ondersteunt voorgaande kritieken met betrekking tot korte termijn effecten van computer gemedieerd leren. Daarenboven toonde deze studie aan dat een debriefing sessie na het spelen van het game geen meerwaarde bood ten opzichte van leerlingen die enkel het game speelden. Dit staat in tegenstelling tot de literatuur waarin wordt gesuggereerd dat een debriefingsessie onmisbaar is bij leren via digitale games.

Dit heeft enkele vragen naar voor gebracht met betrekking tot de aflijning van kenmerken van educatieve videogames die een debriefing sessie noodzakelijk maken. Er is meer bepaald nuancering nodig met betrekking tot game type, complexiteit van de leerinhoud, impliciete/expliciete leerdoelen, game karakteristieken en mogelijks nog andere factoren.

In de tweede haalbaarheidsstudie (hoofdstuk 6) werd aangetoond dat het toedienen van een pre-test in een effectiviteitsstudie de resultaten van de post-test kan vertekenen. Meer bepaald scoorden de deelnemers die een pre-test voor de klassikale les kregen significant hoger op de post-test dan deelnemers die geen pre-test voor de klassikale les kregen, wijzend op pre-test sensitisatie. Bij de deelnemers die de leerinhoud door middel van een game aangeleerd kregen, werden geen significante verschillen gevonden tussen degenen die geen of wel een pre-test kregen voor het spelen van de game. Dit bemoeilijkt de vergelijking tussen een traditionelere les en een game in een pre-test post-test controlegroep experiment, aangezien de scores bij deelnemers in een traditionele les vertekend kunnen zijn. Een mogelijke verklaring voor het feit dat pre-test sensitisatie enkel plaatsvindt in de traditionele lesgroep, is dat de interactiviteit van het game de aandacht en verwerking van de leerinhoud van de deelnemers vergt om verder te kunnen gaan in het spel, onafhankelijk van het feit of ze nu een pre-test kregen of niet. Dit bevestigt de effectiviteit van het game nogmaals. Daarenboven scoorden beide gamegroepen (met en zonder pre-test) beter op de post-test dan de groep die een pre-test kregen voor de traditionele les, wat de effectiviteit van het game nogmaals bevestigt. Deze studie heeft dus ook de meerwaarde van het Solomon 4-groepen design aangetoond, aangezien wij nu beter ondersteunde uitspraken kunnen doen over de effectiviteit van het game.

In onze laatste haalbaarheidsstudie (hoofdstuk 7) die werd uitgevoerd in een bedrijfscontext bleek de interactiviteit van de game geen meerwaarde te bieden ten opzichte van een leervideo die exact dezelfde inhoud aanbood. Het motivatie-idee achter educatieve videogames hield dus geen stand in deze studie, wat er op wijst dat men goed moet afwegen waar interactieve inhoud een meerwaarde kan bieden.

3. Reflectie

Gebaseerd op dit doctoraat, kunnen we stellen dat effectiviteit van educatieve videogames een complex construct is. De verschillende dimensies en sub dimensies die naar voor gebracht werden in onze conceptualisatie, kunnen op twee verschillende manieren benaderd worden. Meer bepaald, kwam het verschil tussen absolute (i.e., het bereiken van vooropgestelde doelen)

en relatieve effectiviteit (i.e., vergelijking van leer-, motivatie- en efficiëntie-uitkomsten met andere leermedia) naar voor. Welk type effectiviteit dient onderzocht te worden, hangt af van de onderzoeksvraag en welke media er momenteel voorhanden zijn om een bepaald onderwerp te behandelen.

Met betrekking tot het meten van effectiviteit, is controle over zo veel mogelijk elementen niet altijd mogelijk noch wenselijk in de context van educatieve videogames. Een eerste reden is de complexe omgeving waar educatieve videogames worden geïntegreerd, zoals implementatie in natuurlijke collectieven (vb. klasgroepen) waar de onderzoeker niet altijd controle heeft over geobserveerde en niet-geobserveerde variabelen. Een andere reden waarom controle niet altijd wenselijk is, is omdat het idee achter het instructiepotentieel van digitale games één van motivatie is. Implementatie in een gecontroleerde lab-omgeving zou hier dus een beperkt inzicht kunnen geven in motivatie-uitkomsten. De complexiteit van effectiviteit van educatieve videogames zorgt er ook voor dat er soms moet ingegeven worden aan experimentele controle. Bijvoorbeeld, het gelijk houden van instructietijd in de experimentele en controlegroep is niet altijd wenselijk. In een context waar leerlingen betaalde werknemers zijn is een gewenste uitkomst van educatieve videogames een vermindering van de leertijd en bijgevolg, een hogere kostenefficiëntie. Het gelijk houden van instructietijd is dus incompatibel met de gewenste efficiënte-uitkomsten van educatieve videogames. In dergelijke gevallen, dient instructietijd als een uitkomst behandeld te worden en dient de focus te liggen op de vraag: leert men door middel van de game minstens even veel of meer in minder tijd in vergelijking met de huidige leermedia?

Om alles samen te vatten, wordt een evenwicht tussen ecologische validiteit en controle best bereikt door eerst en vooral interne validiteit hoog te houden door de invloed van verstorende variabelen te reduceren gedurende de implementatie van de interventie (vb., ondersteuning van de instructeur, geen extra materiaal toevoegen die gerelateerd is aan de leerinhoud). Daarenboven dient men de mogelijke invloed van verstorende factoren gelijk te houden in beide condities (vb. dag en tijd van de interventie, context van implementatie, etc.). Indien er een onevenwicht is met betrekking tot relevante deelnemersvariabelen, kan dit toegevoegd worden aan de analyse om bepaalde verschillen in rekening te brengen (vb. verschillen op de pre-test). Externe validiteit kan gemaximaliseerd worden door vergelijkbaarheid te waarborgen tussen elementen in de studie en hoe het game zou geïmplementeerd worden in de echte wereld (vb. context van implementatie, implementatie in natuurlijke collectieven, aanwezigheid van de gewoonlijke leerkracht, aanbieden van procedurele hulp, etc.).

4. Conclusie

Dit doctoraat heeft aangetoond dat een meer gestandaardiseerde aanpak voor het meten van effectiviteit van educatieve videogames gericht op cognitieve leeruitkomsten niet alleen mogelijk, maar noodzakelijk is. Een meer gestroomlijnde aanpak is nu mogelijk m.b.t. uitkomst maten door middel van onze conceptualisatie en operationalisering van effectiviteit van educatieve videogames. Daarenboven is een gestandaardiseerde aanpak met betrekking tot studie design nu ook mogelijk tot een bepaald punt (vb. pre-test, post-test, follow-up, context die representatief is voor de echte implementatie, procedurele help gedurende de implementatie, etc.). Voor sommige studie design elementen is complete standaardisatie echter niet mogelijk en zal de invulling afhangen van de onderzoeksvraag en de noden en wensen van de stakeholders voor wie de studie wordt uitgevoerd (vb. effectieve of relatieve effectiviteit, afweging tussen ecologische validiteit en experimentele controle, etc.). Daarom is het bij deze studie design elementen vooral belangrijk dat de onderzoekers accuraat rapporteren en dat ze een genuanceerder beeld geven over andere factoren die potentieel invloed kunnen gehad hebben op de uitkomsten. Tot slot is een meer gestandaardiseerde aanpak niet enkel mogelijk, maar ook noodzakelijk om assumpties te kunnen maken over effectgroottes van educatieve videogames, uitspraken over effectiviteit van educatieve videogames op een algemener niveau te kunnen maken en om verder te kunnen onderzoeken welke kenmerken van digitale games kunnen bijdragen aan de effectiviteit van educatieve videogames.

Executive Summary: English

1. Introduction

A large heterogeneity in study designs assessing the effectiveness of digital game-based (DGBL) and questions raised regarding reliability and validity of certain study design characteristics has resulted in a need for a more common methodology. The aim of this Ph.D. project was to develop a standardized procedure for assessing the effectiveness of DGBL, with a primary focus on games that target cognitive learning outcomes (i.e., knowledge transfer and not aimed at skill development or attitudinal/behavioral change). In a first phase, a first version of the procedure was developed. For this purpose, firstly study design characteristics of published DGBL effectiveness studies aimed towards cognitive learning outcomes were mapped by means of a systematic literature review (chapter 2). Secondly, we conceptualized and operationalized effectiveness of DGBL by means of a user requirements analysis among relevant stakeholder groups (chapter 3). Thirdly, we defined best practices for assessing the effectiveness of DGBL by means of expert interviews (chapter 4) in order to finalize the first version of the procedure. In a second phase, we tested the feasibility of the procedure by means of experimental studies using this procedure as a guideline in order to further optimize it. Based on our experiences in the second stage, we have developed a second version of the procedure.

2. Summary of main findings

2.1. *Issues in published effectiveness studies of DGBL*

The systematic literature review has pointed towards three main issues in the field of effectiveness research on DGBL. Firstly, heterogeneity exists among separate studies. The main causes of this heterogeneity significantly impact results. More specifically, the three main issues causing heterogeneity are different activities that are implemented in the control group, the different measures that are used to assess effectiveness and different statistical techniques for quantifying learning outcomes. A second issue with the DGBL effectiveness research field are suboptimal study designs as a result of confounds. More specifically, a) the addition of extra elements to the intervention (i.e., required reading, debriefing session, etc.), b) the presence, type (i.e., familiar vs unfamiliar) and role (i.e., supervision, procedural help or guidance) of the instructor during the intervention and c) practice effects as a result of the same test administered pre- and post-intervention. These elements make it difficult to know

whether the same beneficial effects would have been found without these elements. A third and last issue with DGBL effectiveness research is that of replication of studies. Very little information is given about how the interventions were implemented (e.g., who was present? In which context was the game played? Was gameplay individual? Etc.), how sampling occurred, how similarity was attained between experimental and control group, which tests were implemented and if tests were developed by researchers themselves, how these were developed.

2.2. *Conceptualizing and operationalizing DGBL effectiveness*

Results of the user requirements analysis has shown that effectiveness of DGBL is a multidimensional construct consisting of three categories of desired outcomes: learning, motivational and efficiency outcomes. For every category of outcomes, several indicators can be used. For learning outcomes this can be a) an increased interest in the subject matter, b) performance (i.e., on a knowledge test) and c) transfer. Motivational outcomes can refer to a) creating a more enjoyable learning experience compared to current instructional media or b) making learners more motivated to learn using DGBL. Efficiency outcomes refer to a) reducing time for teaching a certain content matter or b) providing a more cost-effective solution for teaching a content matter to a certain group of learners. Higher motivational and efficiency outcomes are not a stand-alone reason to implement DGBL, but should still be related to similar learning outcomes achieved by more traditional media.

2.3. *Best practices for assessing DGBL effectiveness*

In chapter 3 we have defined best practices for assessing the effectiveness of DGBL, based on semi-structured interviews with experts on intervention research coming from the field of psychology and educational science. In this chapter, we have detected several potential areas for improvement in the field of DGBL effectiveness research: the implementation of the intervention and the methods employed to assess effectiveness. Regarding implementation of both the interventions in the experimental and control group, several practices were defined that are preferably avoided during the intervention in order to reduce confounds (such as guidance by the instructor, extra elements that consist of substantive information) and which elements could be allowed (e.g., procedural help, training session). Moreover, variables on which similarity between experimental and control condition should be attained were determined (e.g., time exposed to intervention, instructor, day of the week). With regard to the methods

dimension, proposed improvements related to assignment of participants to conditions (e.g., variables to take into account when using blocked randomized design), general design (e.g., necessity of a pre-test and control group), test development (e.g., develop and pilot parallel tests) and testing moments (e.g., follow up after minimum 2 weeks). In sum, this chapter provides best practices that cover all aspects of the study design. While several suggestions have previously been made regarding research design of DGBL effectiveness studies these do not cover all aspects of the research design, such as aspects for which similarity between subjects should be attained between experimental and control group, instructor role and implementation of the intervention.

2.4. *Empirical findings on DGBL effectiveness*

Besides testing the feasibility of and optimizing the procedure, the experimental studies also provided us with some insights regarding the effectiveness of DGBL. Our first feasibility study (chapter 5) has shown that while no significant difference could be found between the group that had learned English vocabulary using DGBL and the group that had received a traditional class by the teacher; at the second post-test -three weeks later- the group that had received the traditional class outperformed the group that was instructed by DGBL. Thus, in the longer term the traditional class proved to be more effective. This supports previously made claims on short term effects in computer-based learning. Moreover, a debriefing session did not add value regarding learning and motivational outcomes to the game-only condition. This goes against part of the literature as it has been suggested that a debriefing is indispensable in digital game-based learning. This has raised a number of questions regarding delineation of DGBL characteristics that require a debriefing. More specifically, nuances regarding game type, complexity of learning content, explicit/implicit learning goals, game characteristics and possibly other factors need to be explored.

In our second feasibility study (chapter 6) we found that adding a pre-test to an effectiveness study of DGBL can influence results as pre-test sensitization only occurs in the group that received a slide-based lecture. More specifically, the participants that received a pre-test before the slide-based lecture had significantly higher post-test scores than the slide-based group that did not receive a pre-test before the lecture. In the game group, no significant differences could be found between those participants that received a pre-test before the game-based training and those that did not. This makes comparison of DGBL and more traditional classes in a pre-test post-test control group design rather difficult, as post-test scores of the

traditional class might be positively biased. However, the fact that pre-test sensitization does not occur in the DGBL group also confirms the effectiveness of DGBL, as the interactivity of the game required them to be attentive, regardless of whether they received a pre-test or not before the DGBL intervention. Furthermore, both game groups still outperformed the slide-based group that received a pre-test before the lecture, confirming the higher effectiveness of the game. This study has thus also shown the added value of conducting a Solomon 4-group design in the context of DGBL.

In our third feasibility study (chapter 7), which took place in a corporate context (chapter 7), the interactivity of the game was found not to add value to a passive instructional video that delivers exactly the same content. The motivation rationale behind DGBL did not hold true in this case pointing to the need for careful consideration as to where to use interactive content.

3. Reflection

Based on this dissertation, we can state that effectiveness of DGBL is a complex construct. The several dimensions and sub dimensions defined in chapter 3 can be approached in different ways. An important distinction that comes forward in this dissertation is the difference between absolute effectiveness (i.e., achievement of predefined goals) and relative effectiveness (i.e., comparison of learning, motivational and efficiency outcomes with other instructional media). What type of effectiveness will be required, will ultimately depend on the research question and what type of media are currently available for teaching a certain content matter.

Regarding assessment of DGBL effectiveness, control of as many elements as possible – which is desired in experimental research- can be problematic in the context of DGBL. A first reason for this is the complex environments in which DGBL is often being implemented, such as natural collectives in which one does not always have control over observed and unobserved variables. Another reason why control is not always possible or desirable is that the main rationale behind implementing games as instructional tools is one of motivation. Hence, implementing a game in a controlled lab setting would provide us with limited insight in motivational outcomes as this is a highly artificial environment. The complexity of DGBL effectiveness also results in a trade-off between control and ecological validity. For instance, keeping instructional time equal between experimental and control group is not always desirable. In a context where learners are paid employees, a reduction of training time and as a result, higher cost-efficiency is often a desired outcome. Hence, keeping instructional time equal is incompatible with the efficiency outcomes of DGBL. In such cases, instructional time

should be treated as an outcome and research should focus on investigating whether learners learn as much or more in less time using the game-based method.

To summarize, a balance between ecological validity and control is thus best achieved by firstly increasing internal validity by reducing the influence of confounding variables (i.e., instructor support, extra material during the intervention, etc.) during implementation of the intervention(s) as much as possible and by keeping potential confounds equal in the experimental and control group (e.g., day and time of the intervention, context of implementation, etc.). If there is an imbalance between groups regarding relevant participant variables that might influence the outcomes, this could be added to the analysis in order to take this difference into account (e.g., differences regarding prior knowledge). External validity can be maximized by ensuring similarity between elements present in the real world implementation environment and the implementation for the effectiveness assessment, such as implementation in a context in which the game is intended to be used, implementation in natural collectives such as existing class groups (i.e., randomization on a classroom level or blocked random assignment), the presence of a familiar teacher in a classroom, the provision of procedural help, etcetera.

4. Conclusion

This dissertation has shown that a more standardized approach for assessing DGBL effectiveness is not only possible, but required. Firstly, a more streamlined approach on outcome measures is now possible by our conceptualization and operationalization of DGBL effectiveness. Secondly, a more standardized approach regarding actual study design characteristics is possible up to a certain point (e.g., pre-test, post-test, follow-up, context that is representative for real-world implementation, procedural help during implementation, etc.). However, for some study design aspects, complete standardization is not possible and will ultimately depend on the research question and needs of the people for whom the study is conducted (e.g., assessment of absolute effectiveness, trade-off between ecological validity and experimental control, etc.). Important here is accurate reporting by researchers in order to provide readers with a more nuanced view on factors potentially influencing the outcomes of interest. Finally, a more standardized approach is not only possible, but required in order to make assumptions of the magnitude of the effect of DGBL by means of effect sizes, make claims on a more general level and further investigate features of DGBL that contribute to its effectiveness.

CHAPTER 1.
GENERAL INTRODUCTION

1. Background

In the first world war, propaganda made its appearance in order to influence the public opinion. At the end of the sixties, the TV show Sesame Street used the entertaining power of television to educate. Today, commercials are increasingly interrupting our favorite TV show during primetime in order to influence our purchasing behavior. Media have since their rise believed to have the potential to influence people's attitudes, belief systems, cognition and behavior (McQuail, 2010). Hence, one of the primary focusses of mass communication has been the social, cultural and psychological effects of media contents and its use (Perse. 2001).

Media effects can be subdivided into three categories: cognitive, affective and behavioral. Cognitive media effects refer to the influence media can have on our knowledge and belief system and thus concern media as tools for acquisition of information. Examples of cognitive media effects are, foreign language acquisition as a result of watching subtitled television programs (d'Ydewalle & Van de Poel, 1999) or the impact of women's magazines on thin ideals (Grabe, Ward, & Hyde, 2008). Affective media effects are related to the influence media can have on attitude formation, for instance, positive attitudes towards smoking as a result of tobacco advertising or positive images of tobacco in films (Wellman, Sugarman, DiFranza, & Winickoff, 2006) or gender socialization as a result of stereotypical sex role patterns on television and in films (Smith & Granados, 2009). Affective media effects can also relate to the impact of media on emotions, such as fear or anxiety as a result of scary television (Pearce & Field, 2016) or the usage of Facebook on feelings of loneliness (Song et al., 2014). Behavioral media effects refer to the influence media can have on behavior. This can for instance be exergames influencing children's physical activity (Daley, 2009) or online news use influencing political engagement (Boulianne, 2009).

Media influences can occur on a micro-level, referring to influences on an individual, or on a macro-level, which refers to influences on an audience aggregate, such as institutions, certain societal groups, or on society as a whole (Perse, 2001). A useful example that clarifies the difference between micro- and macro-level media effects is the educational TV show *Sesame Street*, which leads to learning effects among children from all socio-economic classes (micro-level). However, the TV show also widened the knowledge gap between children from high and low socio-economic statuses (SES) as the children coming from higher SES households learned at a faster rate compared to children from lower SES

households (macro-level) (Cook et al., 1975). This example also shows that media effects can differ based on intentionality: media influence can be intended (e.g., learning) or unintended (e.g., widening the knowledge gap). Furthermore, media influences can either be short or long term (McQuail, 2010).

Media influences studied in this dissertation are those of digital games, more specifically the effects of providing instructional content by means of digital games.

1.1. *Digital game-based learning as growing market*

Digital games have become an important part of today's society. In 2015, the global game market's revenue was 90 million dollars (Newzoo, 2015a). In Belgium, 4.2 million people are active gamers (Newzoo, 2015b) and in the USA, half of the Americans adults play videogames (Duggan, 2015). The revenues for educational games are currently estimated at 1.8 billion dollars (Adkins, 2015). The biggest buyers in 2014 can be found in Asia, with a revenue of 1,1123.47 million dollars of which China accounts for 621.7 million; followed by North America (466.13 million) and Western Europe (135.69 million). Consumers account for the largest part of the global revenue (1,362.64 million), followed by corporations (165.73 million) and primary education (143.8 million).

On a policy level, interest in and support for using games as instructional media or Digital Game-Based Learning (DGBL) keeps growing. For instance, in Flanders (Belgium), the Game Fund (part of the Flemish Audiovisual Fund) has been established in 2012, in order to stimulate the gaming industry in Flanders. Two types of funds exist here: a fund for *serious games in compulsory education* (200 000 euros per year) and a fund for *other serious games and entertainment games* (550 000 euros per year). On a European level, there is also a belief in the instructional potential of digital games. For instance, the European Commission invested in research investigating the potential of digital games for empowerment and inclusion (Stewart et al., 2013). Moreover, by means of H2020 funding, development of instructional digital games is stimulated. For instance, in 2016, a call was sent out specifically aimed towards transfer of gaming technologies in non-leisure contexts.

Given this increasing interest and resources that are being allocated for digital game-based learning (DGBL); it is important to know whether the theoretical foundations that underlie digital game-based learning hold true in real life and thus and whether the products that are currently being developed actually succeed in achieving their goals. Meta-analyses have shown mixed results on the effectiveness of DGBL (Backlund & Hendrix, 2013; Clark, Tanner-Smith,

& Killingsworth, 2015; Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012; Wouters, Van Nimwegen, Van Oostendorp, & Van Der Spek, 2013). Moreover, certain authors have pointed out elements that jeopardize reliability and validity of some findings (Clark, 2007; Clark et al., 2015; Hays, 2005), stating that positive results might not have been found when a more rigorous design would have been applied.

The aim of this dissertation was to develop a standardized procedure for assessing the effectiveness of DGBL (with a primary focus on games that target cognitive learning outcomes). In this introductory chapter, we will first discuss the relationship of media and learning. Secondly, we will define digital game-based learning and its position in the field of entertainment education. Thirdly, we will discuss motivational and instructional theoretical underpinnings as to why digital games can be a successful instructional medium. Fourthly, we will point towards some issues regarding the validation of the effectiveness of DGBL. Lastly, we will describe the research objectives of this dissertation and methodology used to reach these objectives.

2. Media & learning

Since the early days of mass communication research, it has been common to study the differential effectiveness of media for information processing and learning (Valkenburg, Peter, & Walther, 2016). In the literature, however, there has been discussion as to whether media can actually influence learning. This is often referred to as 'The great media debate' between Richard Clark and Robert Kozma. Since the introduction of computers to education until current implementations of games in education, Clark has stated that media can never influence learning, but that they are mere vehicles through which instruction is delivered. The effects found are thus a result of the instructional methods embedded in the media presentation. Hence, the choice of media is irrelevant to the effectiveness of instruction and choices for a certain medium to deliver a certain instructional method should be based on efficiency and cost (Clark, 1994). According to Kozma (1994), however, the medium and learning content are inherently connected. More specifically, different media consist of different attributes or capabilities that can influence the effectiveness of instruction. Hence, both the media and the instructional method are part of the instructional design:

> In good designs, a medium's capabilities enable methods and the methods that are used take advantage of those capabilities. If media are going to influence learning, method must be confounded with the medium. Media must be designed to give us powerful new methods, and our methods must take appropriate advantage of a medium's capabilities
> (Kozma, 1994, p. 2).

In line with Kozma (1994), we consider that learning is indeed a result of the method embedded in a medium but that certain media possess certain attributes that allow for certain methods to be embedded (i.e., interactivity). This can possibly result in a more effective instructional tool. On the other hand, we also agree with Clark (1994), that decisions to implement a certain medium as instructional tool can also merely be based on efficiency and cost, without the medium providing an added value regarding learning outcomes. However, the main rationale behind implementing a certain medium for instruction can also be related to its motivational power –which in turn can influence learning- that again is a result of attributes certain media types possess, which again is in line with Kozma's vision.

3. Digital game-based learning

One field that relies on the assumption that media do influence learning is Digital Game-Based Learning (DGBL). Digital games encompass a variety of types and genres that can be played using a multitude of digital technologies such as computers, (handheld) consoles and mobile devices. Based on a literature review of digital games definitions, Juul (2003) defines a digital game as

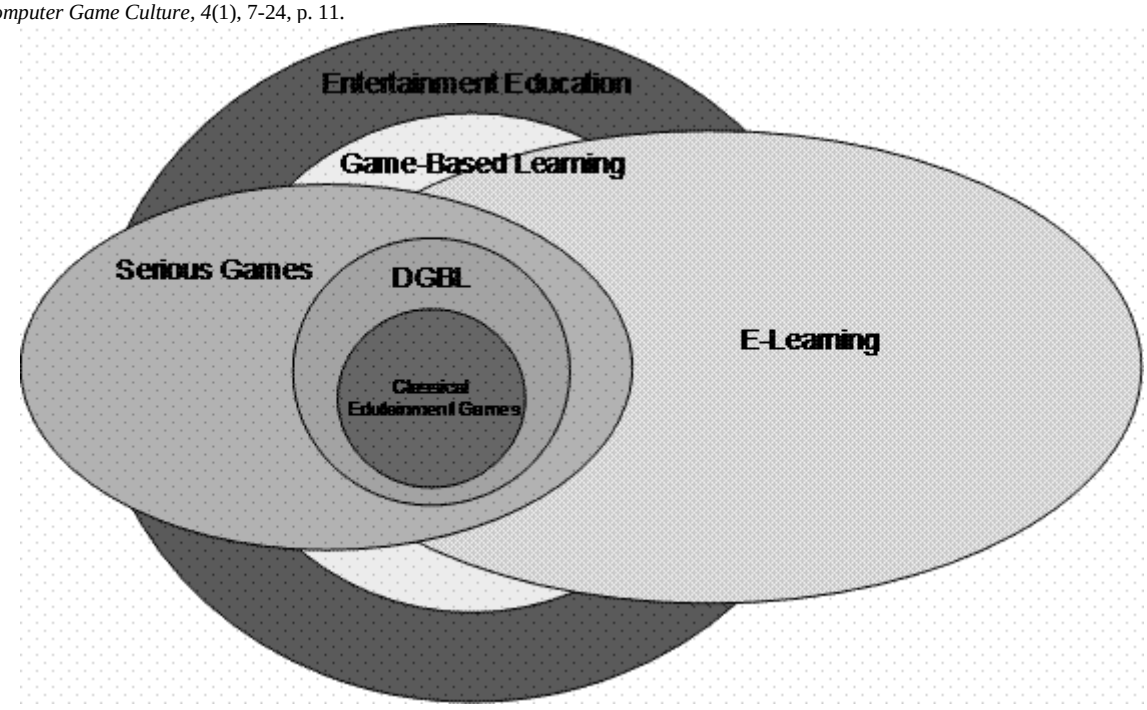
> …a rule-based formal system with a variable and quantifiable outcome, where different outcomes are assigned different values, the player exerts effort in order to influence the outcome, the player feels attached to the outcome, and the consequences of the activity are optional and negotiable (p.5).

DGBL refers to the usage of the entertaining power of digital games to serve an educational purpose (Prensky, 2001). The goal of DGBL is thus twofold: it has to be fun/entertaining and it has to be educational (Bellotti et al. 2013). Before we discuss why digital games are considered a good medium for instruction, we will firstly introduce related terms to position

digital-game based learning and why we did not choose other similar concepts, such as serious games. Digital game-based learning can be considered a form of **edutainment** or **entertainment education**, referring to the usage of entertainment media such as television to serve an educational purpose (Susi, Johannesson, & Backlund, 2007). An example of edutainment is the children's TV show *Sesame Street*. DGBL can also be considered a subcategory of **e-learning**, which refers to "…technology-based learning in which learning materials are delivered electronically to remote learners via a computer network" (Zhang, Zhao, Zhou, & Nunamaker Jr, 2004, p. 2). DGBL can also be considered a subcategory of **serious games**, which is a more broader term that has applications outside the field of education, such as games for health, therapy or games made with job recruitment purposes. **Game-based learning** (without the qualifier 'digital') is also a broader term and can also refer to learning by means of board games. **Classical edutainment games** refer to a series of educational games that have made their appearance in the 1990s (e.g., Where in the world is Carmen Sandiego?). Lastly, **Classical entertainment** games can also be used for educational purposes. This is often referred to as Commercial of the Shelf games or COTS (Stewart et al., 2013). Figure 1 provides a schematic overview of Digital Game-Based Learning and similar instructional media concepts (Breuer & Bente, 2010).

Fig. 1: Relationship between DGBL and similar entertainment education concepts

Taken from Breuer, J. S., & Bente, G. (2010). Why so serious? On the relation of serious games and learning. *Eludamos. Journal for Computer Game Culture*, *4*(1), 7-24, p. 11.



There are several reasons why scholars believe that digital games are considered an appropriate medium for instruction. First, digital games possess certain attributes that can positively influence the learner's motivation to start and persist in the educational intervention. Secondly, games also contain certain attributes that allow the implementation of certain learning paradigms. Below, we discuss the main rationales for implementation of games as instructional tools for their motivational (section 3.1.) and instructional power (section 3.2.).

3.1. *Motivational benefits of digital game-based learning.*

3.1.1. Intrinsic motivation

The power of games to intrinsically motivate players to engage in the activity has been considered as an important characteristic which can benefit learning (Garris, Ahlers, & Driskell, 2002). Intrinsic motivation is a concept from self-determination theory (Ryan & Deci, 2000a, Ryan & Deci, 2000b) and distinguishes between intrinsic motivation and extrinsic motivation. Intrinsic motivation refers to doing an activity in itself for itself, because it is inherently enjoyable and/or interesting. Those activities have an appeal of curiosity, novelty or aesthetic for individuals. Extrinsic motivation refers to doing an activity, because it will lead to a separable outcome (Ryan & Deci, 2000a). The concept of intrinsic motivation is especially relevant for digital game-based learning as intrinsic motivation to perform an activity is associated with higher levels of enjoyment, interest, performance, higher quality of learning and heightened self-esteem (Ryan & Deci, 2000b).

Intrinsic motivation, however, is often assumed in the context of gaming, but is not always a reality. Especially in the context of DGBL, players can be primarily extrinsically motivated to participate, referring to engaging in the activity as a result of external coercion, influencing enjoyment of the activity and consequently, learning outcomes (Boyle et al., 2011; Mayer et al., 2014). However, intrinsic motivation is not a perquisite for learning and several types of extrinsic motivation can also have a positive impact on learning. According to self-determination theory (Ryan & Deci, 2000a) extrinsic motivation can be nuanced and subdivided in different types, depending on the extent to which their regulation is internalized (i.e., inner acceptance of the value or utility of an activity) or are thus more autonomous. Higher levels of autonomy of regulation related to extrinsic motivation result in higher levels of engagement (Connell & Wellborn, 1991), performance (Miserandino, 1996), higher quality of learning (Grolnick & Ryan, 1987) and lower levels of dropout (Vallerand & Reid, 1984).

Internalization of the regulation related to the motivation can be stimulated by a sense of relatedness (i.e., a sense of belongingness and connectedness to the persons, group, or culture disseminating a goal), competence (i.e., self-efficacy) and autonomy support (i.e., create a sense of volition) (Ryan & Deci, 2000b). The least autonomous form of extrinsic motivation is *external regulation*, which refers to an activity that is performed in order to receive a reward or avoid some negative contingency. An example of external regulation is engaging in DGBL in order to receive extra credits for a class. A more autonomous form of extrinsic motivation is *introjected regulation* and refers to an activity that is performed out of a sense of guilt or obligation or a need to prove something. Engaging in DGBL in a classroom context out of fear of being evaluated negatively by the teacher is an example of introjected regulation. The second most autonomous form of extrinsic motivation is *identified regulation* which refers to the performance of an activity, because the action or the outcome is accepted as personally important. An example of this type of regulation is engaging in DGBL for programming, because it will help the player to achieve his goal of becoming a programmer. *Integrated regulation* is the most autonomous type of external motivation and refers to regulations that are fully assimilated to the self and are consistent with other goals and values. When a pupil engages in DGBL in a school context, because he/she wants to be a good student, is an example of integrated regulation.

Shifts from external to more internalized types of motivation or even to intrinsic motivation in a learning context can be stimulated by characteristics of the learning environment (Ryan & Deci, 2000a; Ryan & Deci, 2000b). Table 1 provides a non-exhaustive overview of game features that can foster intrinsic motivation.

Table 1: Game features that can foster intrinsic motivation

Character-/player related	Game related	Related to graphical representation
Autonomy/sense of control Triggered by sense of choice, character management (Clark, 2007; Dickey 2007, Hsu et al, 2005; Malone, 1981)	Challenge (Banarowski, 2007; Clark, 2007, Hsu et al, 2005)	Realism Characters/world are similar to the real world representations (Hsu et al, 2005)
Self-efficacy Triggered by balance between challenge and skill-level; positive feedback (Clark, 2007; Malone, 1981)	Integrated goals (Clark, 2007)	Fantasy Stimulation of imagination (Banarowksi, 2007; Malone, 1981)
Curiosity Triggered by quests, hidden information (Malone, 1981)	Feedback (Clark, 2007; Banarowski, 2007)	
Identification With in-game character and game world (Hsu et al, 2005)	Interactivity (Banarowski, 2007; Clark, 2007, Hsu et al, 2005)	
Sense of powerfulness Triggerd by powerful weapons or skills, can in turn influence self-efficacy (Hsu et al, 2005; Banarowksi, 2007)	Rewards (Banarwoski, 2007; Hsu et al, 2005)	
	Narrative (Hsu et al, 2005)	
	Novelty (Hsu et al, 2005)	
	Cooperation (Dickey, 2007)	

3.1.2. Flow theory

Flow is a concept that refers to the mental state where concentration is so intense that one loses notion of the self and of time (Csikszentmihalyi, 1990). This mental state is reached when an activity is performed which in itself is satisfying and enjoyable; the activity must thus be intrinsically motivating. Characteristics of activities that can lead to this flow state are related to the ability to complete the task, the ability to concentrate on the task as a result of clear goals and immediate feedback and sense of control over actions (Csikzentmihalyi, 1990). Examples of activities that can produce a flow experience are dancing, making music, exercising, but also playing video games (Chiang, Sunny, Chao-Yang, & Liu, 2011; Csikszentmihalyi, 2013; Hoffman & Novak, 2009). A prerequisite to attain a flow state is to achieve a balance between a person's (perceived) skills to perform an activity and the challenges that are associated with executing that activity. Both should attain a certain level to result in a flow state. When there is no balance between the challenge related to tasks in an activity and skills to perform these tasks, a flow state will not be achieved. When someone, for instance, learns a new sport, skills will rather be low while challenge will be high. The fear to fail will thus be somewhat higher. As

the skill level improves, however, and the challenge remains high by learning new tasks or subcomponents of that activity, anxiety will develop into arousal. When skill level in turn, becomes more advanced, a state of flow can be reached (Csikszentmihlayi, 1990).

Flow theory is a framework that is often referred to when explaining why games create an enjoying experience as game play experiences are consistent with dimensions of flow experience (Kiili, de Freitas, Arnab, & Lainema, 2012). In order to create an optimal game experience -similar to antecedents for creating a flow experience- clear goals, feedback, sense of control and an appropriate balance between challenge and skills should be integrated in a digital game (Kiili et al., 2012; Sweetser & Wyeth, 2005). More specifically, the **overriding goal** of the game should be made clear from the beginning and intermediate goals should be presented at appropriate times (Federoff, 2002; Pagulayan et al., 2002). **Feedback** can consist of several elements: progress towards the goals, immediate feedback on in-game actions and the ability to know your status or score in the game at any given time (Sweetser & Wyeth, 2005). **Challenge** is considered the most important aspect of a good game design. Difficulty levels of a video game should be variable to meet all players at the correct level of challenge. To create this balance between challenge and **skill level**, games typically start out with a beginners level which gradually increases in difficulty as the player's skills progress (Desurvire, Caplan, & Toth, 2004; Pagulayan et al., 2002). A sense of **control** – for instance, on the movements, character and/or interface- is also an important factor in creating an enjoyable game experience: the player must feel like his actions have an impact on the game world and the game world should react to the players' actions (Desurvire et al., 2004). The player should have the feeling that he is able to do what he wants and not have the feeling that he is working through a path fixed by the game developer.

One known consequence of flow experience in computer mediated environments are increased learning (Skadberg & Kimmel, 2004). A flow state can enhance the learning process due to the intense concentration during the activity (Korteling, Helsdingen, Sluimer, Emmerik, & Kappé, 2011). Moreover, the balance between skill level and challenge is in line with Vygotsky's *zone of proximal development*, according to which learning content is best processed when the challenge is at that difficulty level that the learner can manage successfully (Boyle, Connolly, & Hainey, 2011; Ritterfeld & Weber, 2006). Also, while in *flow* state, the learner is completely motivated to push his/her skills to the limit, which is a highly desirable state in an instructional context (Paras, 2005).

3.1.3. Entertainment-Education paradigm

While indeed games can be enjoyable and intrinsically motivating, providing educational content in a game format will not automatically result in learners wanting to start an instruction just because it is a game. Ritterfeld and Weber's (2006) paradigm for entertainment education provides a less strong, more nuanced vision on how the entertaining value of digital games can be used for educational purposes.

Ritterfeld and Weber (2006) propose that the relationship between entertainment and education can either be **linear positive**, referring to entertainment as a facilitator for learning content; **linear negative**, where entertainment distracts from learning and results in a decrease of the learning performance or **inverse U-shaped**, meaning that entertainment can positively influence learning, but only until a certain point, after which it is detrimental for the learning outcome (Ritterfeld & Weber, 2006).

Three paradigms can be distinguished on how digital games can be implemented in education. In the **motivation paradigm**, entertainment is a facilitator for learning content and games are implemented to 'seduce' the learners to allocate their attention to the learning content. The main focus here, is fostering motivation for continuation and consequently, learning (Ritterfeld, Weber, Fernandes, & Vorderer, 2004). Providing a narrative around certain science exercises is an example of the motivation paradigm in DGBL. In the **reinforcement paradigm**, the entertaining parts of games are provided as a reward for learning. Here, the aim is to use entertainment to foster extrinsic motivation for processing educational content. Examples of reinforcement strategies in DGBL are scores, virtual money, fun animations, or the reward of progress in the video game play (Ritterfeld & Weber, 2006). In both the motivation and reinforcement paradigm, learning goals are explicit and intentional. Most games developed for DGBL follow the motivation and/or reinforcement paradigm to combine entertainment with education (Ritterfeld & Weber, 2006). In the **blending** paradigm, entertainment and learning content are intertwined: the enjoyment of mastery in the game is equivalent to the enjoyment of the acquisition and use of knowledge and skills. Learning here is implicit and incidental (Ritterfeld, Weber, Fernandes, & Vorderer, 2004). This can for instance refer to accidentally learning some history facts as the result of playing a history game.

3.2. *Instructional benefits of DGBL*

Digital games allow for the implementation of constructivist theories of learning (Boyle et al., 2011; Rooney, 2012). Constructivism relies on the assumption that learning is a process in which learners' knowledge and skills are *constructed* by making sense of their experiences. In constructivist learning theory, the learner is an active learner as opposed to a passive one receiving and processing information provided by an instructor (Hein, 1991). Main constructivist learning mechanisms that underpin the instructional potential of digital game-based learning are situated learning, experiential learning and problem-based learning (Boyle et al., 2011; Rooney, 2012). Games can enable **situated learning**, according to which learning is context-dependent and needs to occur in the context of the authentic learning environment to which the learning applies (environment, actions, situations and actors) (Ladley, 2010). An authentic learning environment is one that replicates what the learner would experience in a real-world situation. Learning is thus a result of the interaction of mental processes with the physical and social environment (Clancey, 1991). In certain cases such as emergency situations, a simulation of that authentic environment is the best alternative solution for providing this situated learning experience (Ladley, 2010). Digital games have the ability to provide this authentic environment, both regarding the simulation of the actual physical environment, events and consequences of actions made in this simulated world.

Digital games also enable an **experiential learning experience**, according to which experiences are a source of learning and one learns by doing (Kolb, 1984). According to Kolb, an experiential learning experience is a cyclical process which consists of four phases. The first phase is the *concrete experience*, followed by the second phase, *reflective observations*, where the learner observes and reflects on this experience. Based on these observations and reflections, the learner draws conclusions and makes hypotheses and generalizations on how this acquired knowledge can be used in other situations, which is called *abstract conceptualization*. The final phase in this cyclical process is *active experimentation*, where the learner tests these hypotheses by experimenting and applying the acquired knowledge. This process also occurs while playing video games, requiring **"…**a constant cycle of hypothesis formulation, testing, and revision. This process happens rapidly while the game is played, with immediate feedback" (Van Eck, 2006, p. 5).

Digital games also offer the potential to provide a **problem-based learning** experience (Van Eck, 2015), where a particular problem is presented to the learners and knowledge and

skills are acquired during the process of solving this problem (Savery & Duffy, 1995). Problem solving is a mechanism that often occurs in digital games, by means of goals or missions a player has to accomplish (Kiili, 2005).

4. Empirical evidence for Digital Game-Based Learning

The increased interest in and implementation of DGBL has resulted in a need to know whether or not these instructional tools are effective (Mayer et al., 2014) and whether the substantial financial effort required to develop and implement DGBL is worthwhile (Clark, 2007). Although single case studies and meta-analyses have proven the effectiveness of DGBL (Backlund & Hendrix, 2013; 2015; Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012), results are still mixed and the current state of the art does not allow us to conclude that educational games and simulations have a positive effect on learning and motivation (Erhel & Jamet, 2013; Giessen, 2015). Certain authors have pointed out elements that jeopardize reliability and validity of some findings (Clark, 2007; Clark, Tanner-Smith, & Killingsworth, 2015). This includes comparisons with control groups that did not receive an educational intervention (Hays, 2005), time-on-task differences between experimental and control groups, and validity of research instruments (Randel, Morris, Wetzel, & Whitehill, 1992). Moreover, some studies do not provide enough information about the implementation of the intervention (Clark et al., 2015; Sitzmann, 2011). This makes it hard for readers to know if the reported results are a consequence of the different methods, and not a cause of circumstantial factors that differed between conditions (Randel et al., 1992). Rigorous assessment is required to improve the quality of DGBL, to support resource allocation, and to gain insight in the most effective way to use games to support learning (De Freitas, 2006; Kirriemuir & McFarlane, 2004).

Moreover, there is a large heterogeneity in study designs used to assess the effectiveness of DGBL. For example, different research designs are applied, different measures are used for assessing effectiveness and different statistical techniques are used to quantify learning outcomes (All, Nuñez Castellar, & Van Looy, 2014; Kharrazi, Lu, Gharghabi, & Coleman, 2012). An underlying reason for this is that DGBL is an emerging field, which combines different disciplines with specific research traditions (Kirriemuir & McFarlane, 2004; Van Eck, 2015). Hence, there is a need for an overarching methodology to research and evaluate DGBL, which should provide procedures, frameworks, and methods that can be validated (Mayer et al., 2014). While several suggestions have been made to improve the design of

DGBL effectiveness studies (Brom et al., 2012; Serrano-Laguna et al., 2013), these do not cover all aspects of the experimental research design (e.g., aspects for which similarity between subjects should be attained, instructor role).

A common methodology would firstly create the opportunity to compare results, and thus the outcomes, of the different instructional media and methods across studies. Secondly, claims regarding the effectiveness of DGBL could be made on a more generalized level: per field (e.g., science, math, language learning) or per game genre. Thirdly, a common methodology would set a baseline for quality, which could serve as an evaluation tool for published studies and as a starting point for researchers wanting to conduct a DGBL effectiveness study. Lastly, interest in studying game design features (e.g., competition, feedback, narrative) that influence effectiveness is growing in order to optimize DGBL game design. In order to make general claims about game design features that influence effectiveness (i.e., by means of meta-analyses such as conducted by Desmet and colleagues in 2014), a more standardized approach for studying effectiveness is required (Cagiltay, Ozcelik, & Ozcelik, 2015; Kirriemuir & McFarlane, 2004), as a prerequisite for meta-analyses is indeed that there is some homogeneity of the research design dimensions of the studies involved (Higgins, Green, & Collaboration, 2008).

A final reason why some guidelines or best practices should be defined in the field of digital game-based learning is that there is a publication bias in media effects research in general (Perse, 2001) and effectiveness research on digital game-based learning (Chiu, Kao, & Reynolds, 2012). Hence, studies reporting significant results or large effect sizes are more likely to be published. However, it is also important to know when effects based on literature are not what they are expected to be in order to know whether it is worth making the investment. Consequently, if studies produce null findings, but are conducted according to certain validated guidelines, this might increase publication opportunities of these null findings.

5. Research objectives

The aim of this Ph.D. project is to develop a procedure for assessing the effectiveness of digital game-based learning by means of experimental research. We have chosen experimental

research because this dissertation is situated at a crossroad of media-effects research and evaluations of educational interventions.

5.1. *Development of a procedure using experimental methodology*

Although both qualitative and quantitative methods are used for the study of media-effects, it has a strong quantitative focus centered around causal inference (Nabi & Oliver, 2009; Perse, 2001). Qualitative research in the study of media-effects is more focused on audience reception studies. More specifically, the subject of study here is the audience as an agent of exposure to certain media content instead of audience as an outcome. Research questions in qualitative media-effects research relate to for instance meanings people attribute to media content, identity formation and media as a tool for interpersonal relationships (Nabi & Oliver, 2009). The focus of this dissertation is, however, on audience as an outcome and will thus develop a procedure rooted in quantitative methodology.

Common quantitative methodologies for studying media effects are survey and experimental research (Nabi & Oliver, 2009; Wimmer & Dominick, 2013). In survey research, typically measures of media-exposure are self-report measures of the outcome of interest are assessed (e.g., anti-social behavior as a result of heavily consuming television). In experimental research, this exposure is manipulated in a more controlled setting, followed by the assessment of the outcome of interest. In this dissertation, the aim is to establish whether or not the implementation of instructional content in digital games can be beneficial for learning and will thus typically require a manipulation, more specifically implementation of a digital game-based learning intervention. Moreover, an experimental approach is considered most suitable as this dissertation can be positioned in the field of evaluation of educational interventions as well, where two types of evaluation of educational interventions can be distinguished (Calder, 2013). A first type is formative evaluation which aims to determine areas for improvement and is thus an evaluation of the process of the intervention itself. This type of evaluation is conducted by using a naturalistic design with observational data collection, which describes an ongoing process in its natural setting. A second type is summative evaluation, which aims at determinining whether or not an educational intervention succeeds in attaining its goals, thus evaluating the outcomes or its effectiveness (Calder, 2013). Summative evaluations are conducted by using an experimental design (Hutchinson, 1999). In the present dissertation, we focus on summative evaluations (i.e., determining its effectiveness) and will concordantly focus on the development of a procedure using an experimental design.

5.2. *Focus on DGBL that primarily aimed at cognitive learning outcomes*

Based on the projected primary learning outcomes, three types of DGBL can be distinguished aiming at knowledge transfer (cognitive learning outcomes), skill acquisition (skill-based learning outcomes) or attitudinal/ behavioral change (affective learning outcomes) (Stewart et al., 2013). Games that primarily aim at knowledge transfer are typically implemented in education, in order to teach math (Nunez Castellar, Van Looy, Szmalec, & De Marez, 2013) or language (Yip & Kwan, 2006) for example. Digital games that primarily aim at skill acquisition are used for training, for example in a corporate or military context. Several such studies have examined the impact of playing games to practice managerial skills (Corsi et al., 2006; Kretschmann, 2012). Games aimed at attitudinal change are also used by governments and NGOs to raise awareness of a certain topic such as poverty (Neys, Van Looy, De Grove, & Jansz, 2012). Games aimed at behavioral change are typically found in the health sector, for example for promoting healthier eating or physical activity among children (Baranowski, Buday, Thompson, & Baranowski, 2008). Learning is, however, a multidimensional construct and while DGBL can primarily aim at a certain type of learning outcome, it can entail secondary learning outcomes (Kraiger, Ford, & Salas, 1993). For instance, a game that primarily aims at teaching children English (cognitive learning outcomes) can also result in a more positive attitude towards learning English or English as a subject (affective learning outcomes).

In the present doctoral dissertation we only focus on the development of a procedure for assessing the effectiveness of DGBL that primarily aims towards cognitive learning outcomes. We do not include DGBL that primarily aim towards skill development or attitudinal/behavioral change, considering that different types of learning outcomes require different types of assessment and thus resist categorization in one research taxonomy (Kraiger, Ford, & Salas, 1993). Cognition refers to knowledge and ideas/beliefs of an individual and the mental activity regarding processes such as learning, thinking, interpreting and problem-solving. Cognitive learning outcomes consists of three categories: verbal/encoded knowledge, knowledge organization and cognitive strategies. *Verbal/encoded knowledge* refers to knowledge that is transferred through the written or spoken word (Gagne, 1984; Kraiger et al., 1993). Verbal knowledge can refer to *declarative knowledge* (i.e., information about what), *procedural knowledge* (i.e., information about how) and *strategic knowledge* (i.e., information about which, when and why). *Knowledge organization* refers to the way knowledge is structured in the mind and is also referred to as mental models. When for instance, a problem needs to be

solved, an individual makes use of mental models that show similarities with the problem he/she is facing in order to solve it (Kraiger, Ford, & Salas, 1993). Finally, *cognitive strategies* refer to the usage of personal strategies to learn, think, act and feel (Gagne, 1984).

5.3. *Research objectives*

To develop this procedure, four research objectives were created in this Ph.D. A first research objective was to map how DGBL effectiveness studies are currently being conducted. A second research objective was to conceptualize and operationalize DGBL effectiveness, as there is currently no clear evaluation framework available. A third research objective was to define best practices regarding effectiveness assessment of DGBL. A fourth research objective was to optimize the guidelines by means of 3 feasibility studies in sectors where DGBL are currently being implemented: school education, the health sector and corporate education.

6. Methodology

Three phases are conceived in this Ph.D. project: the development of the procedure version 1.0, feasability studies and the development of the final procedure.

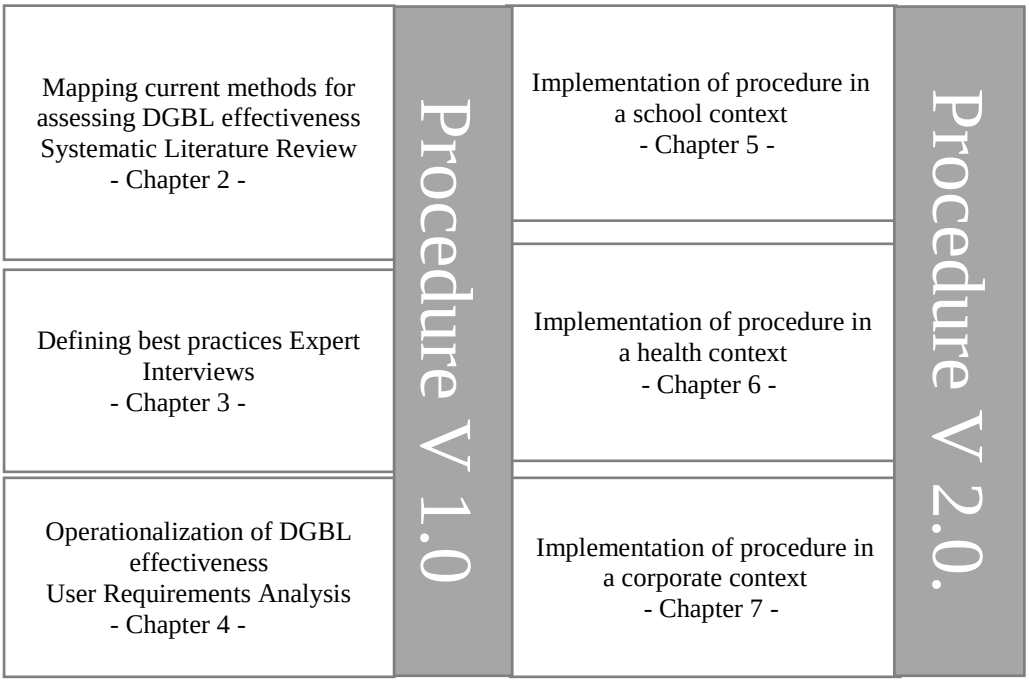
The first phase consists of three subphases. In the first subphase, the methods that are currently being used for sampling, implementation of the interventions, measures and data analysis were mapped in a systematic literature review using Cochrane guidelines. The results of this can be found in chapter 2. In a second subphase, a conceptual framework for assessing the effectiveness of DGBL was developed based on a user requirements analysis with (potential) adopters of DGBL (containing business), game developers and DGBL researchers (operational working area) and stakeholders on a governmental level (wider environment). For this purpose, three focus groups were conducted; one for each stakeholder group. The results of this can be found in chapter 3. In the third subphase, the variety in study characteristics brought forward in the systematic literature review (subphase 1) were presented to experts in psychology and pedagogy by means of semi-structured interviews, in order to define best practices for conducting DGBL effectiveness studies. The results of this can be found in chapter 4. Based on the studies conducted in phase 1, a first version of the procedure was developed.

In the second phase of this project, the procedure was used as a guideline in three effectiveness studies in the three main areas where DGBL is currently being implemented: a school context (chapter 5), a health context (chapter 6) and a corporate context (chapter 7) in

order to test the feasibility of the procedure. This way, the best practices can be optimized in order to develop a standardized procedure for assessing the effectiveness of DGBL that can be flexibly used across different sectors.

In a third phase, a final version of the procedure was developed based on the results and experiences of the validation studies conducted in the second phase of this project. The reflections of implementation of the procedure in the different contexts can be found in the final chapter of this dissertation (chapter 8). A description of the procedure can be found in appendix. The actual procedure can be retrieved from the authors upon request.

Fig. 3: Schematic overview of dissertation



PART 1.

TOWARDS THE DEVELOPMENT OF GUIDELINES FOR ASSESSING THE EFFECTIVENESS OF DIGITAL GAME-BASED LEARNING AIMED TOWARDS COGNITIVE LEARNING OUTCOMES

CHAPTER 2.
Measuring Effectiveness in Digital Game-Based Learning: A Methodological Review.

Abstract

In recent years, a growing number of studies are being conducted into the effectiveness of digital game-based learning (DGBL). Despite this growing interest, there is a lack of sound empirical evidence on the effectiveness of DGBL due to different outcome measures for assessing effectiveness, varying methods of data collection and inconclusive or difficult to interpret results. This has resulted in a need for an overarching methodology for assessing the effectiveness of DGBL. The present study took a first step in this direction by mapping current methods used for assessing the effectiveness of DGBL. Results showed that currently, comparison of results across studies and thus looking at effectiveness of DGBL on a more general level is problematic due to diversity in and suboptimal study designs. Variety in study design relates to three issues, namely different activities that are implemented in the control groups, different measures for assessing the effectiveness of DGBL and the use of different statistical techniques for analyzing learning outcomes. Suboptimal study designs are the result of variables confounding study results. Possible confounds that were brought forward in this review are elements that are added to the game as part of the educational intervention (e.g., required reading, debriefing session), instructor influences and practice effects when using the same test pre- and post-intervention. Lastly, incomplete information on the study design impedes replication of studies and thus falsification of study results.

Keywords*:*

Digital Game-Based Learning, Effectiveness assessment, cognitive learning outcomes

Reference:

All, A., Nunez Castellar, E. P., & Van Looy, J. (2014). Measuring effectiveness in digital game-based learning: a methodological review. *International Journal of Serious Games*, 2(1), 3–20.

1. Introduction

In recent years, there has been a growing interest in the potential of games as instructional tools in areas such as education, health and wellbeing, government, NGOs, corporate, defense, marketing and communication (Sawyer & Smith, 2008). Considering that the development and implementation of digital game-based learning (DGBL) implies a substantial financial effort, there is an increasing need to determine the educational potential of DGBL in order to justify the investment (Clark, 2007; Mayer, Bekebrede, Warmelink & Zhou, 2013). One major justification of this investment should be well-founded empirical evidence (Clark, 2007). While in recent years, there has been an increasing number of publications aimed at assessing the effectiveness of DGBL, there is still a lack of sound empirical evidence (Mayer, 2011). The lack of an overarching methodology for effectiveness research on DGBL has led to the use of different outcome measures for assessing effectiveness (O'Neil, Wainess & Baker, 2005), varying methods of data collection (Ke, 2009) and inconclusive or difficult to interpret results (Clark, 2007). Moreover, questions have been raised regarding the validity of current effectiveness research on DGBL (Clark, 2007; O'Neil, Wainess & Baker, 2005; Hays, 2005). A common methodology for assessing the effectiveness of DGBL would firstly create the opportunity to compare results and thus the quality of the different educational interventions across studies. Secondly, claims regarding the effectiveness of DGBL could be made on a more general level. Lastly, a common methodology could set a baseline for quality, which could serve as an evaluation tool for published studies and as a starting point for researchers desiring to conduct an effectiveness study on DGBL. The present study aims at mapping current research methods used for effectiveness research on DGBL and is a first part of a larger project aimed at the development of a standardized procedure for assessing the effectiveness of DGBL.

1.1. *Defining effectiveness of DGBL*

Based on the projected primary learning outcomes, three types of DGBL can be distinguished aiming at knowledge transfer (cognitive learning outcomes), skill acquisition (skill-based learning outcomes) or attitudinal/ behavioral change (affective learning outcomes) (Stewart et al., 2012). Games that primarily aim at knowledge transfer are typically implemented in education, in order to teach math (Nuñez Castellar, Szmalec, Van Looy & De Marez) or language (Yip & Kwan, 2006) for example. Digital games that primarily aim at skill acquisition

are used for training, for example in a corporate or military context. Several studies have for instance examined the impact of playing games to practice managerial skills (Corsi et al., 2006; Kretschmann, 2012). Games aimed at attitudinal change are also used by governments and NGOs to raise awareness of a certain topic such as poverty (Neys, Van Looy, De Grove & Jansz, 2012). Games aimed at behavioral change are typically found in the health sector, for example games promoting healthier food and physical activity to children (Baranowski, Buday, Thompson & Baranowski, 2008). Learning is, however, a multidimensional construct and while DGBL can primarily aim at a certain type of learning outcome, it can entail secondary learning outcomes. For instance, a game that primarily aims at teaching children English (cognitive learning outcomes) can also result in a more positive attitude towards learning English or English as a subject (affective learning outcomes).

According to O'Neill et al. (2005) effectiveness of DGBL can be defined in terms of 1) intensity and longevity of engagement with a game 2) commercial success of a game and 3) acquisition of knowledge and skills as a result of the implementation of a game as an instructional medium. In the current study, we will focus on the third aspect, and more specifically on the acquisition of knowledge.

The effectiveness of DGBL as an instructional medium firstly consists of first order learning effects, referring to a direct influence on knowledge, skills, attitudes or behavior. This is typically assessed by looking at changes between pre- and post-game measurements (Mayer, Bekebrede, Warmelink & Zhou, 2013). A second aspect of effectiveness of DGBL is transfer, referring to the application of the learning content to real world situations (Korteling, Helsdingen, Sluimer, Emmerik & Kappé, 2011). This is typically assessed gathering data in the field, such as key performance indicators or by organizing a follow-up test (Mayer, Bekebrede, Warmelink & Zhou, 2013). As mentioned before, primary learning outcomes can entail certain secondary learning outcomes (e.g., a game that aims at teaching math skills can also lead to a more positive attitude towards math). In the case of educational interventions, especially when choosing for DGBL, motivation is often a secondary learning outcome one wishes to attain. Motivation is a necessary prerequisite to ensure that learners actually learn something. When they are not motivated, the chance of failing of an educational program will increase (Gunter, Kenny & Vick, 2006). Moreover, according to Kozma (1994) medium and learning content are inherently connected, implying that characteristics of the medium can influence the learning outcome. The power of games to intrinsically motivate players to engage in the activity (i.e.,

performing the activity in itself and for itself (Ryan & Deci, 2000a)) has been considered as an important aspect of games which can benefit learning (Garris & Driskell, 2000). More specifically, intrinsic motivation for performing an activity is associated with higher levels of enjoyment, interest, performance, higher quality of learning and a heightened self-esteem (Ryan & Deci, 2000b). This type of motivation, however, is often assumed in the context of gaming, but is not always a reality. Especially in the context of DGBL, players can be extrinsically motivated to participate, referring to engaging in the activity as a result of external coercion. However, extrinsic motivation can be nuanced and subdivided in different types, depending on the extent to which their regulation is autonomous. The least autonomous form of extrinsic motivation is external regulation, which refers to an activity that is performed in order to receive a reward or avoid some negative contingency. An example of external regulation is engaging in DGBL in order to receive extra credits for a certain class. A more autonomous form of extrinsic motivation is introjected regulation and refers to an activity that is performed out of a sense of guilt or obligation or a need to prove something. Engaging in DGBL in a classroom context out of fear of negatively being evaluated by the teacher is an example of introjected regulation. The second most autonomous form of extrinsic motivation is identified regulation which refers to the performing of an activity, because the action or the outcome is accepted as personally important. An example of this type of regulation is engaging in DGBL for programming, because it will help the player to achieve his goal of becoming a programmer. Integrated regulation is the most autonomous type of external motivation and refers to regulations that are fully assimilated to the self and are consistent with other goals and values. For instance, when a pupil engages in DGBL in a school context, because he/she wants to be a good student, is an example of integrated regulation. These different types of extrinsic motivation are also associated with different outcomes and experiences. More specifically, higher levels of autonomy of extrinsic motivation result in higher levels of engagement, performance, higher quality of learning and lower levels of dropout. How autonomous the external motivation is, depends on the level of internalization of regulations or values. Internalization of regulation and values can, however, be stimulated by the feeling of relatedness with significant others modeling or valuing a certain behavior. Perceived competence (i.e. self-efficacy) and the experience of autonomy (i.e. feeling of volition) (Ryan & Deci, 2000b) also play an important role in this internalization process.

2. Evaluation of educational interventions

Educational evaluation aims at describing and explaining experiences of students and teachers and judging the effectiveness of education (wilkes & Bligh, 1999). Two types of evaluation can be distinguished: formative and summative evaluation. Formative evaluation aims at detecting areas for improvement, thus evaluating the process, whereas summative evaluation aims at determining to what extent an educational intervention was successful, thus judging its effectiveness (Calder, 2003). While summative evaluation can occur independently, formative evaluation cannot occur without a summative evaluation (Taras, 2005).

Educational evaluation is not the same as educational research which requires more rigorous standards of reliability and validity (Hutchinson, 1999). Educational research can be conducted in two ways: by using a naturalistic design, describing an ongoing process in its natural setting, mostly by using observations or by using an experimental design which evaluates the impact of an educational intervention on its desired learning outcomes. DGBL effectiveness research should thus strive for more rigorous standards of validity and reliability in order to be considered as educational research, which underlines the need for defining standards.

3. DGBL effectiveness studies

The most implemented designs in DGBL effectiveness studies are quasi-experimental and survey design. A study of Chen and O'Neill (2005) has shown that in most empirical studies on DGBL effectiveness, no pre-test of knowledge is implemented. According to Clark (2007) the absence of a pre-test of knowledge is problematic, because differences in learning outcomes could be due to knowledge differences between individuals or groups at the start of the intervention. Consequently, this can lead to an overestimation of the instructional effect.

Moreover, when control groups are included in the studies, often no educational activity is implemented in the control group (O'Neil, Wainess & Baker, 2005; Hays, 2005). According to Hays (2005) the comparison to a control group, which does not receive an intervention or does not engage in educational exercises, is problematic in this type of research because, again, it might lead to an overestimation of the beneficial effects of DGBL. This is also supported by Clark (2007) who states that one of the major motivations for the use of DGBL should be the justification of the investment made and should thus be compared to viable and less expensive

alternative ways to teach the same knowledge and skills. According to Clark (2007), this comparison should also be made on motivational aspects, and more specifically on motivation to learn through the game-based approach compared to other instructional programs.

Questionnaires are typically used to assess the motivational aspects of DGBL, gauging the motivations of participants for learning via the intervention received and their interest in participation (Hainey, 2006). Questions have been raised by several authors in the field about the validity of these measures (Wouters, van der Spek & Van Oostendorp, 2009) considering student opinion on for example learning and motivation has previously been found to be unreliable and conflicting with direct measures (Clark, 2007). Suggestions have been made towards physiological or behavioral measures (e.g., eye-tracking, skin conductance), because data can be collected during game play in a more controlled manner (Wouters, van der Spek & Van Oostendorp, 2009). Furthermore, motivation as a construct in the context of DGBL effectiveness research needs to be further examined since questions can be raised on whether definitions of motivation in different studies truly represent motivation or other constructs (Wouters, van der Spek & Van Oostendorp, 2009). Further, questionnaires are also implemented to assess other affective outcomes, such as attitudes (Wouters, van der Spek & Van Oostendorp, 2009).

Some studies use in-game assessment – referred to as stealth assessment – which is a technique that aims at accurately and dynamically measuring the player's progress, analyzing the player's competencies at various levels during game play (Shute, Rieber & Van Eck, 2011). Using technology, which strategies the player uses to solve certain problems can for instance be assessed in the game, giving the researcher information on the learner's progress (Shute, 2011). Finally, qualitative methods such as interviews (e.g., attitudes before game play, player experiences after game play) and observation (e.g., behavioural performance after playing game, decision making and emotional reactions during game play) have also been used in the context of effectiveness studies of DGBL (Mayer, Bekebrede, Warmelink & Zhou, 2013).

4. Method

In the present study the Cochrane method was used to carry out our systematic literature review (Higgins, 2008). This review method has its origins in health research and aims to study the effectiveness of interventions for prevention, treatment and rehabilitation. According to Cochrane, for dimensions of study characteristics can be distinguished: 1) participants (e.g., characteristics of the sample involved), 2) intervention (e.g., contents, format, timings and

treatment lengths, intervention(s) in control group(s)), 3) methods (e.g., applied research methods) and 4) outcome measures (e.g., instruments used to measure a certain outcome) and results.

For this review, we only included studies that implemented games which primarily aim at cognitive learning outcomes, considering the different types of learning outcomes require different types of assessment and thus resist categorization in one research taxonomy (Kraiger, Ford & Salas, 1993).

Search engines used for our review were Web of Knowledge, EBSCO Host and the International Bibliography of the Social Sciences. The following search string was used: "((Edu* OR serious OR learn* OR digital game based learning) AND ((dig* OR video OR computer) AND game) AND (assess* OR effect* OR measur*))". This search identified 54 publications dealing with effectiveness of DGBL aimed at cognitive learning outcomes. Criteria for inclusion were that (1) the publications were peer-reviewed journal and conference publications between 2000 and 2012 (2) the focus was on digital games and (2) a pre-post design with a control group was used. According to Campbell and Stanley (1963), a pre-post control group design is the best design to assess learning considering that a pre-test offers the opportunity to measure progress and a control group ensures us that this progress is not due to a mere lapse of time. Eight studies had a post-only design with a control group and 21 studies had a pre-post design without a control group which were all excluded. Eventually, 25 studies with a pre-post design and control group were considered eligible for analysis.

A quantitative content analysis was conducted using SPSS. The codebook for this analysis was created by coding the methods and procedures sections in the studies both deductively (fixed dimensions of study design based on Cochrane) and inductively (methods and elements belonging to dimensions of the study design) in nVivo. Open coding was used for identifying different methods and creating labels (e.g., randomization of subjects, randomization of classrooms, matching of participants). Subsequently, axial coding was used for creating categories by relating labels to each other representing different elements of the study design (e.g., assignment of participants). Lastly, the categories were assigned to the different dimensions of the study design as defined by Cochrane.

5. Results

5.1. *Participants*

Inclusion criteria for participation in the studies were mostly school-related (e.g., 'majoring in math and science). Other studies included a certain subgroup, including participants based on ability (e.g., low achievers), socioeconomic status (SES) or a certain health condition.

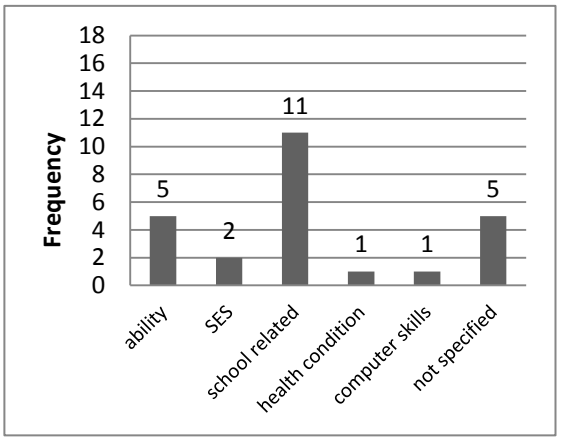


Figure 1: Inclusion criteria (N = 25)

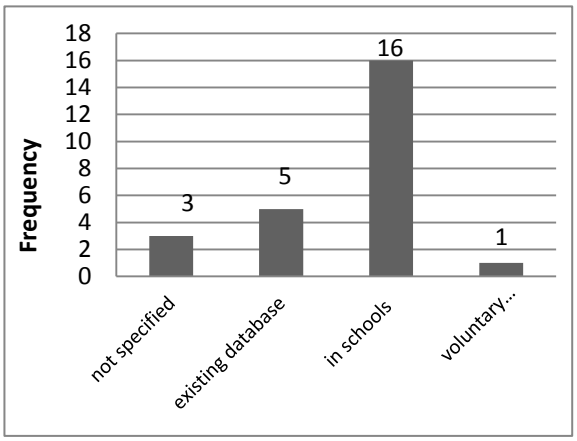


Figure 2: Recruitment of participants (N = 25)

Twenty per cent did not specify inclusion criteria used for participants (Fig. 1). Participants were mostly recruited in schools and by using existing databases. One study recruited based on voluntary participation and 3 studies did not specify how participants were recruited (Fig. 2).

The average sample size of participants in studies reviewed was 220 (SD = 284, Mdn = 100), with a minimum of 6 and a maximum of 1274 participants. A minimum of 6 participants spread over several conditions is a remarkably low sample size when carrying out statistical analyses and making certain claims regarding the effectiveness of that particular game, let alone generalizing claims on DGBL effectiveness based on the results of that particular study. Although not all the studies reported the number of participants included by group (8% did not), our results showed that when reported the average number of participants was 105 (SD = 163, Mdn = 46), with a minimum of 2 and a maximum of 758 participants in the experimental and 84 (SD = 92, Mdn = 45) with a minimum of 2 and a maximum of 347 in the control group. Although four studies reported participants' mean age, most studies

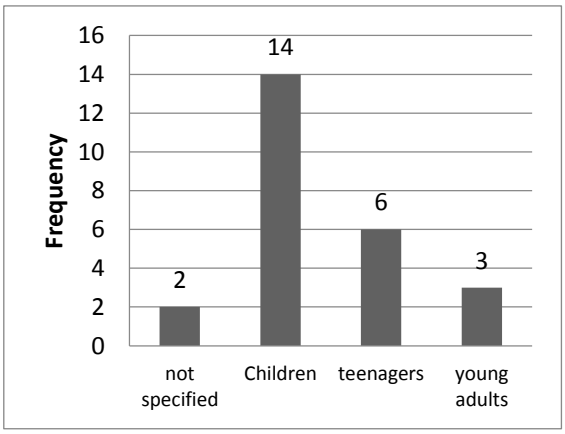


Figure 3: Subjects included in the study (N = 25)

defined subjects based on types of people, such as 'university students'. Sixty-five per cent of the studies included children, 24% teenagers and 12% young adults (Fig. 3).

5.2.*Intervention*

In the majority of the studies (64%) DGBL was implemented in a formal context (e.g., in school during school hours), 8% in an informal context (e.g., home setting) and 12% in a semi-formal context referring to an implementation in a formal institution, such as a school, but where gameplay occurred outside of school hours (Fig. 4). Sixteen per cent did not specify the context of play and 56% did not specify the gameplay composition. Twenty-four per cent let participants play individually, 4% individually in competition, 24% cooperatively and 4% in a cooperative competition, meaning groups of participants played together against other groups of participants. One study implemented all for gameplay conditions (Fig. 5). Results of the latter study showed that game play composition influences learning outcomes. More specifically, individual gameplay leads to a significantly better performance. Therefore, 56% studies failing to report on game play composition is problematic.

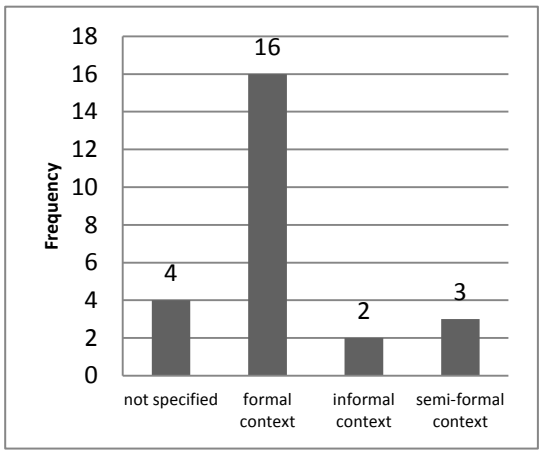


Figure 4: Context of gameplay (N = 25)

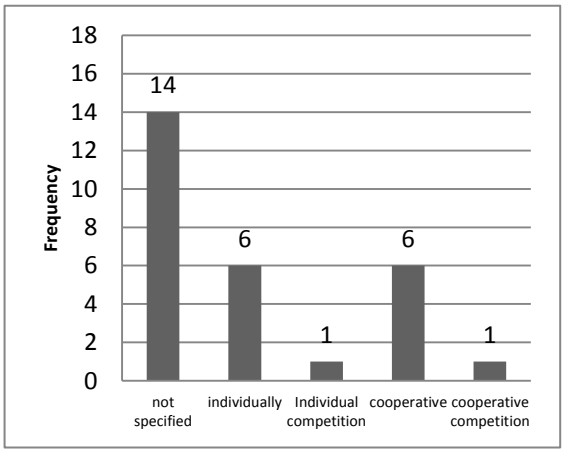


Figure 5: Gameplay composition (N = 25)

Games were either implemented as a stand-alone intervention (28%) or were embedded in a larger program (48%). Forty per cent of the studies did not report on the presence of an instructor, referring to a teacher or researcher present during gameplay. In 56% of the studies an instructor was reported to be present: 5 studies included a teacher as an instructor, 4 studies a researcher, 2 studies university students and 3 studies did not specify the type of instructor present during gameplay. One study did not include any instructor. Several studies implemented the game as a supplement of a course. However, half of these provided extra time for the experimental group to interact with the game in addition to the courses thus spending additional

time with the learning content, leading to confounding effects. Twenty-four per cent did not specify implementation. Table 1 shows an overview of program specifications. While it could be beneficial for DGBL to add elements to the intervention in order to enhance its effectiveness, for the purpose of research aiming at examining whether or not a specific game is effective leads to certain issues. More specifically, these could lead to confounding effects making it impossible for the researcher to know if the positive effects in favor of DGBL were the result of the game as such or the combination of the game with other elements. This is especially problematic when elements containing substantive information regarding the learning content of the game (e.g. extra material, required reading) are added to the DGBL intervention.

Table 1: Specifications of games embedded in a larger program (N = 11)

Program specifications	N	%	Description	Examples from studies reviewed
Introduction	5	20	An introduction concerning game content and gameplay was provided by an instructor. This does not refer to an in-game introduction	*…basic instruction in the area of daily economics…Next, the students were shown how to play the game in order to achieve the stated objectives* [24a]
Training of participants before intervention	5	20	A training session before the intervention was provided	*…children were introduced to a 'treasure hunt game' to allow them to develop the skills necessary to navigate in the virtual world of the computer* [3a]
Extra material	8	32	Extra material such as articles, extra exercises, extra reading material, etc. were freely available	*…two classroom instructors, the study guide, their fellow classmates, referenced publications,…* [16a]
Online platform	3	12	The game was part of a larger educational online platform	*Two vocabulary web sites …. Vocabulary games were also available* [25a]
Game task formulation	1	4	Certain tasks were formulated during gameplay	*The students worked together to play the game and synthesize their answers* [24a]
Required reading	2	5	The participants were expected to read next to gameplay	*…required reading for the students were the lab documents…* [1a]
Procedural help by instructor	3	12	The participants received help concerning the actual gameplay. This does not relate to content	*The Computer Science teachers were present in order to provide procedural help to the students, without, however, being actively involved* [15a]
Guidance by instructor	3	12	The participants received guidance during gameplay in order to contextualize the game in the broader learning context	*…instructional discussion between the students and the teacher while the students were playing the game* [5a]
Supplement of course	6	24	Gameplay occurred next to the classes	*After teaching to both groups all required concepts in a regular classroom… a regular set of exercises was given as homework for two weeks to the students of both groups, while students from the test group, apart from the regular exercises interacted with the game during same period of time* [13a]
Debriefing	3	12	A debriefing session was provided	*Once a play event finished, the instructor held a traditional 45-minute 'discussion section' with the students* [17a]

The average implementation period was 9 weeks (SD = 6,7, Mdn = 6), with a minimum of 1 day and a maximum of 23 weeks. Average total interaction time with the game is 12.4 hours (SD = 14.8, M = 9), with a minimum of 30 minutes and a maximum of 64 hours.

Experimental groups (EG) were compared to a control group (CG) that either included participants that did not get an intervention (24%), got an intervention using another instructional approach (56%), or were compared to several control groups, combining both (16%). One study did not provide any information on interventions implemented in the CG. Table 2 gives an overview of interventions implemented in the control group(s). Thirty-two per cent of the studies reported on how similarity of content in the intervention in the EG and CG was achieved, 24% did not report on this and 12% used dissimilar interventions regarding content. The latter is problematic, considering that in order to make claims on the added value of the DGBL intervention, it should be compared to another educational intervention, covering the same content and preferably instructional techniques with the digital game aspect being the only difference.

Table 2: Interventions in control group (N = 25)

Intervention control group(s)	**N**	**%**	**Description**
Traditional classroom teaching	12	48	(One of) the control group(s) got a comparable treatment/intervention by classical classroom teaching
Traditional multimedia classroom teaching	1	4	(One of) the control group(s) got a comparable treatment/intervention by classical classroom teaching with the help of multimedia (video, audio, etc.)
Computer-based learning	4	16	(One of) the control group(s) got a comparable treatment/intervention by a computer-based application, such as an educational website.
Other game	2	8	(One of) the control group(s) got a treatment/intervention by means of another game not related to the subject concerning the game played in the intervention group
Paper and pencil exercises	3	12	One of the control group(s) got a comparable treatment/intervention by means of paper-and-pencil exercises.
No intervention	10	40	(One of) the control group(s) did not get a comparable interventions, bur served as a no-treatment control group.
Not specified	1	4	The study did not report on the type of intervention implemented in the control group(s)

5.3.*Method*

All studies implemented a quantitative research approach, 32% combined this with qualitative research such as observation, interviews and diaries. However, only 3 studies coded their qualitative data.

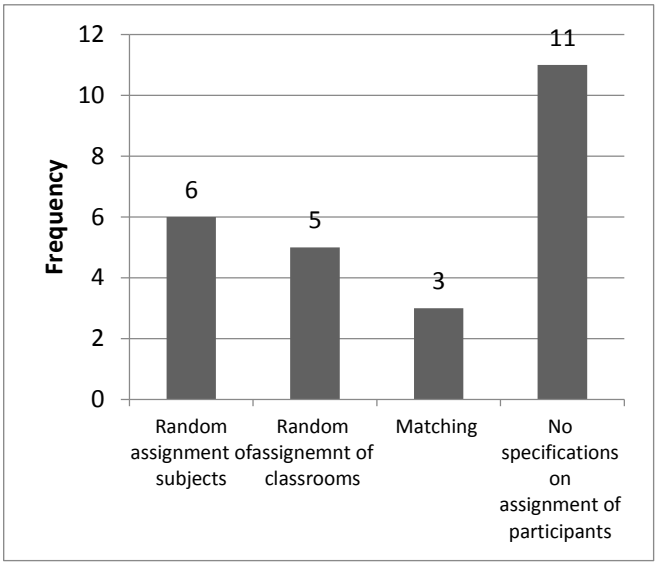


Figure 6: Assignment of participants to conditions

All studies reviewed implemented an experimental design. All studies implemented a between-subjects design, with the exception of one study that implemented a within-subjects design, where the game-based group also served as a control group (by implementing traditional classroom teaching before midterm exams and implementing the DGBL intervention before the final exams). Forty-four per cent used a randomized controlled trial; 24% randomly assigned subjects while 20% randomly assigned classrooms to one of the conditions. Twelve per cent did not randomly assign participants to experimental and control group(s), but 'matched' participants in groups based on certain characteristics such as previous test scores, and 44% did not specify on group assignment of participants (Fig. 6).

5.4.*Measures*

Twenty per cent of the studies reviewed only implemented tests developed by the researchers and 24% used school tests or exams ('student achievement') as an accuracy measure. Two studies (8%) used both test scores and student achievement as an accuracy measure. Less than half (44%) implemented standardized tests, six of these (55%) only used standardized tests while 5 studies (45%) combined standardized tests with tests developed by the researchers. Table 3 gives an overview of measures used in the studies. Thirty-six per cent of the studies reported on how scoring on tests occurred. Three studies (12%) included an independent coder, of which two controlled for inter-rater reliability. One study used several, non-independent coders to control for inter-rater reliability.

Table 3 (part 1): Measures used for determining effectiveness (N = 25)

Objective measurements	**N**	**%**	**Description**	
Accuracy	**19**	**76**		
Test scores	16	64	Absolute test scores of a test developed for the study or a standardized test that has been implemented in the study	*The Civics and Society test (CST) was developed using materials provided by the textbook publisher* [33]
Student achievement	5	24	Student achievement in the formal context (e.g., exam scores)	*…the outcome performance on the midterm examination served as a comparison matched-control, while the outcome performance on the final examination represented the post–digital game-based examination test group.* [6a]
Time measurements	**2**	**8**		
Time on task	2	8	Time spent on finishing tests	*Time taken to complete the challenge was recorded.* [34]
Subjective measurements	**N**	**%**	**Description**	
Self-measurements	**8**	**32**		
Self-efficacy topic	4	16	Self-efficacy concerning the topic of the game	*I'm confident I can understand the basic concepts taught in this course, I believe I will receive an excellent grade in this class* [33]
Self-efficacy general	2	8	Self-efficacy on a more general level (e.g., academic achievement)	*General academic self [35]*
Perceived educational value	2	8	Perceived educational value of the intervention	*…questionnaires…in the experimental group in order to evaluate the online resources in terms of their design and effectiveness in helping them learn vocabulary [11]*
Motivation	**10**	**40**		
Motivation towards educational intervention - Post-only, EG - Post-only, EG and CG - Pre- and post, EG and CG	7 3 2 2	28 12 8 8	Motivation towards learning via a certain intervention	*the degree to which they found that the application: (1) was interesting, (2) was enjoyable, (3) was engaging* [15a]
Motivation towards learning/educational content - Post-only, EG and CG - Pre-post, EG and CG	3 2 1	12	Motivation towards the actual educational content and not to the way it was delivered	*Motivated Strategies for Learning Questionnaire* [24a]

Table 3 (part 2): Measures used for determining effectiveness (N = 25)

Other	2	8		
Attitudes towards school	1	4	Measures for attitudes towards school	*…instrument designed to assess children's attitudes toward primary school* [12a]
Teacher expectations	1	4	Teachers' expectation of change	*In the pretest, teachers must indicate changes expected…In the post-test, teachers must identify positive and negative changes perceived in the dimensions indicated in the pretest…* [19a]

Twenty-eight per cent did not report on the similarity between the pre- and post-test measurements. Forty per cent employed the same test before and after the intervention, 8% changed the sequence of the questions and 8% used a similar test (e.g., other questions with the same type and difficulty levels). The latter did not report on how similarity of parallel tests was assessed. Sixteen per cent used a dissimilar pre- and post-test, such as midterm exam scores and final exam scores. Two studies also implemented a mid-test and for studies a follow-up test. Assessing the lasting effect is, however, important considering that short-term interventions with a new medium can yield a novelty effect, overestimating the instructional value.

Different statistical techniques can be distinguished for quantifying learning outcomes. The larger part of the studies (76%) did a check on pre-existing differences between experimental and control group(s) and 36% of the studies included in this review reported on effect size. Table 4 shows how analysis of tests occurred.

Table 4: Data-analysis (N = 25)

Data-analysis	Description	N	%	Examples from studies reviewed
Between groups comparison of difference scores	The difference (e.g., gain scores or percentage of improvement) between pre- and post-test scores are calculated and used as dependent variable in a between groups comparison (e.g., anova, t-test)	9	36	*…paired-samples t tests were conducted to compare the treatment and control gain scores from pre-test to post-test…[8a]*
Absolute test scores comparison	Differences between experimental and control group are calculated separately for the pre-test (controlling for pre-existing differences) and the post-test scores (e.g., anova, t-test).	5	20	*…the independent samples t-test was applied to examine whether the differences between the mean scores of the control and experimental groups in the pre-test and post-test were statistically significant [25a]*
Pre-test scores as covariate between subjects	Between groups comparison of absolute post-test scores, controlling for initial levels of ability adding pre-test scores as a covariate	4	16	*A 2 x 2 between-groups analysis of covariance (ANCOVA) was conducted to assess the effectiveness of the interventions on students' computer memory knowledge. The independent variables were: (a) the type of intervention, which included two levels (gaming application, non-gaming application), and (b) gender. The dependent variable consisted of scores on the post-test CMKT. Students' scores on the pre-test CMKT served as a covariate in this analysis, to control for eventual pre-existing differences between the groups [15a]*
Between groups comparison with repeated measures	Interaction between time (pre-test and post-test) and group (EG and CG) are calculated (e.g., mixed Anova)	4	16	*The NTPS scores were analyzed using a two-way mixed design ANOVA, in which instructional treatment was a between-subject factor, while measurement occasion was a within-subject factor [24a]*
Repeated measures within subjects	A repeated measures for pre-test and post-test score are calculated separately for experimental and control group(s)	1	4	*Significant gains were found in the games console group for both accuracy and speed of calculations, while results for the two comparison groups were mixed…The comparison groups showed in significant gains in any area of self-perceptions [11a]*
Other	Within subjects design: testing whether or not increased upward shift of scores on pre- and post-tests is statistically significant	1	4	*Though the means and the highest scores remained similar, the lowest score shifted from 53.06% on midterm examination to 57.84% on final examination (post–digital game based outcome). This increased positive upward shift was statistically significant at P 5 .04 [6a]*
Not specified	Results are discussed without describing the data-analysis methods	1	4	/

6. Conclusion

Table 5 gives an overview of the main differences across studies regarding study design. These elements could serve as a foundation for the development of an overarching methodology for assessing effectiveness of DGBL, examining which elements and which ways of execution lead to more reliable and generalizing results on DGBL effectiveness.

Table 5: Summary of main differences across studies

Aspect of study design	Main differences across studies (N=25)
Participants	Large variety in sample size
	Reporting on types of people included
Intervention	Activity implemented in control group(s)
	Stand-alone intervention vs. embedment in a larger program
	Variety of elements present in larger program
	Presence of / role of / type of intermediary
Method	Randomization of subjects/classrooms
	Use of matching in different ways for assigning participants to conditions
	Addition of qualitative data
Measures	Different objective measures of performance
	Different self-report measures
	Similarity pre- and post-tests
	Data-analysis techniques

7. Discussion

The present study indicates that comparison of the results of studies and the making of generalizing claims on DGBL effectiveness is difficult as a result of diversity in study designs, some of which are suboptimal.

Variety in study design is a result of three issues. A first issue is that different activities are implemented in the control group(s). The interpretation of the contribution of the intervention to the EG does, however, depend on the activities performed in the CG (Stewart et al., 2012). Considering that intervention in the CG can influence results and interventions implemented in CG differed across studies, comparison between study results becomes problematic. A second issue regarding variety in study designs is the different measures that are used for assessing effectiveness. While motivation is considered as an important element in DGBL effectiveness, it is not always assessed. When motivation is assessed, the type of motivation measured and timings of measurement differed across studies. The first type of motivation is motivation toward the educational intervention, gauging for engagement and/or enjoyment during game play and is thus a situational component. This is typically related to measuring concepts as enjoyment, fun and immersion. This is, however, somewhat problematic considering this often implies that the motivation for playing games in the context of DGBL is personal motivation or motivation enabled by the game. As mentioned before, engaging in DGBL is mostly the result of external coercion. To become engaged, a player thus firstly needs to be motivated. In turn, to experience enjoyment and immersion, the player needs to be engaged (Schønau-Fog, H. & Bjørner, 2012). A suggestion made by Schønau-Fog and Bjørner (2012) in that respect is assessing the desire to continue playing, investigating the basal level of engagement. The second type of motivation, motivation towards learning or the educational content, however, is seen as an outcome of the intervention. Therefore, it would be interesting to use this measure as a proxy for effectiveness of the educational intervention, considering this could point to a higher interest in the content matter. Consequently, a combination of both types of motivation would be recommended. The development of a validated scale for assessing these types of motivation is therefore an interesting venue for further research.

A third issue is that different statistical techniques are used for quantifying learning outcomes, either comparing gain scores of EG and CG, comparing post-test scores of both groups using pre-test scores as a covariate or using a mixed design, looking at the interaction of time (pre- and post-test) and group (EG, CG). Other studies only compared post-scores, after checking

whether the EG differed significantly from the CG on the pre-test. There has been previous discussion in the academic field on how to analyze data of a pre-post control group design (Singer & Willet, 2003). While the use of gain scores has been criticized as being less reliable than using raw scores, it can be used under certain conditions (i.e., pre-test and post-test scores do not have equal variances and equal reliability). These scores cannot, however, be correlated with other variables in the sample. A mixed design would lead to the same results as comparing gain scores (Dimitrov & Rumrill, 2003). According to several authors, an analysis of covariance (ANCOVA) with pre-test scores as a covariate, is a more preferable method (Campbell & Stanley, 1963; Dimitrov & Rumrill, 2003). In the context of randomized controlled trials, ANCOVA reduces error variance and in the context of nonrandomized designs, it adjusts mean scores of the post-test to differences between groups on pre-test scores (Dimitrov & Rumrill, 2003).

Suboptimal study designs are a result of confounding variables influencing the results, leading to insecurity about whether or not the effects found can be attributed to the game-based intervention or other elements added to the intervention during implementation. Confounds should therefore be eliminated as much as possible (Leary, 1995). There are three types of confounding elements that can be distinguished in the DGBL study design. A first possible confound is the addition of elements to the game used for the intervention. The DGBL intervention is either implemented as a stand-alone intervention or is embedded in a larger program. When embedded in a program, elements of the program differed across studies as well (e.g., introduction, debriefing, extra material, required reading, etc.). The researcher can therefore not know if positive findings are the result of playing the game or the combination of the game with for instance exercises in a textbook, unless this is added as an additional condition to the study (e.g., game, game + textbook, control). A second possible confound is the presence of an instructor. If an instructor was present, the type of instructor (i.e., researcher, teacher, student) and the role of the instructor (i.e., supervision, procedural help, guidance) differed across studies as well. Having a teacher as an instructor in a study can, however, result in less control and as a result, confounding variables (Brom, Šisler, Buchtová, Klement & Levčík, 2012; Serrano-Laguna et al., 2013). Moreover, the presence of an instructor can lead to instructor influences. For instance, a study conducted by Brom et al. (2013) has shown that significant findings in one experimental group compared to its matched control group could not be found in another experimental group compared to its matched control group due to teacher influences. Further, offering procedural help or guidance can again lead to an overestimation of the instructional effect of the DGBL intervention (Joy & Garcia, 2000). A third possible

confound are practice effects when the same test is implemented pre- and post-intervention. when taking an achievement/intelligence test for the second time, participants will automatically do better, even if the intervention would not have taken place. According to Crawford et al. (1989) this is due to retention of specific test material by the participants. Other studies used similar tests, meaning these consisted of questions of the same type and difficulty level. While practice effects can still occur using a parallel version of a test at different points in time (e.g., pre- and post-test), these generally tend to be smaller (Anastasi, 1961). The studies in the review that used parallel tests pre- and post-intervention did not specify how this similarity was assessed however. An example on how this could be done, can be found in a study conducted by Nuñez Castellar et al. (2013) for instance, where similarity of two parallel versions of a test is assessed by providing one half of the participants with version A and the other half with parallel version B in the pre-test and vice versa. Non-significant differences on the pre-test then refer to comparability of version A and B. Other studies also used dissimilar tests, when for example student achievement in school (e.g., exam scores) was used as a measure. This seems problematic, considering assumptions on the comparability of both tests cannot be made, making any significant achievement gains possibly invalid.

Lastly, there are also replication issues with certain studies due to missing information on multiple areas of the study. A detailed description of the procedure is necessary in order to provide other researchers the opportunity to falsify obtained results (Popper, 2000). Most information is missing on implementation of the intervention, sampling, similarity of the different interventions when other educational interventions are implemented in the control group(s) and information on the tests implemented. The latter two also bring doubt to the validity of certain study results. Information on how similarity between different conditions is attained, is necessary for the reader of an academic publication to know whether different groups were treated in the exact same way with the manipulation (e.g., DGBL intervention) being the only difference considering this is a prerequisite for making conclusions on the effect of the manipulation (Leary, 1995). Creating comparable conditions is, however, a challenge considering that comparing interactive media content in a game with for instance an oral class given by a teacher is difficult. A suggestion made by Clark (2007) in that respect is the implementation of similar instructional techniques (e.g. drill and practice, scaffolding) in the control condition. Consequently, differences in learning outcomes can be attributed to the added value of the medium.

Missing information on the tests that are implemented is also problematic, considering that a general problem in this research area seems to be that test development does not always happen

thoroughly enough, again raising questions on their validity (Brom, Šisler, Buchtová, Klement & Levčík, 2012; Serrano-Laguna et al., 2013). When a test is developed by the researchers, little information is provided on the instruments. For instance, there is often no information on whether or not these tests were piloted. This is important information to provide, however, considering educational research requires rigorous standards of reliability and validity, implying that tests developed by researchers should be piloted and include checks on their internal consistency (Hutchinson, 1999). Further, objective tests, subjective tests or a combination of both are used for assessing learning outcomes. The mere use of subjective tests such as self-efficacy is considered as problematic, considering student opinion on learning has previously been found to be unreliable and conflicting with direct measures, questioning their validity (Clark, 2007).

8. Limitations and future research

The selection and coding of publications was conducted by one researcher, which can be considered a limitation of this study. This study is also limited to digital games aimed at cognitive learning outcomes. Further research should thus be conducted on methodologies used in digital games aimed at skill acquisition and behavioral or attitudinal change.

An interesting area for future research is exploring the possibilities for the development of an overarching methodology to measure effectiveness of DGBL. Further research should therefore firstly focus on the development of an evaluation framework for assessing effectiveness of DGBL in order to develop a common methodology. To be able to develop this evaluation framework, a clear definition of effectiveness in the context of DGBL should be formulated. Considering that there are a lot of stakeholders involved in this field (e.g., game designers, game researchers, adopters and governmental institutions providing funding), this definition should not solely be based on literature reviews, but should also include the conceptualization of effectiveness by these different stakeholders. Moreover, both relevant stakeholders and experts in the methodology field (i.e., educational research and experimental methodology) should be involved in the development of a common methodology in order to find a balance between an ideal research design in terms of validity and what is practically possible.

Lastly, some issues have been raised on confounding elements by implementing the game in a larger program. Empirical evidence on the possible impact of these elements in the context of DGBL research is, to the best of our knowledge, scarce. Therefore, further research on the

impact of several factors such as support by intermediaries, program elements and extra material provided, is required.

9. Appendix: Studies included in literature review.

[1a] Anderson, J. and Barnett, M. 2010. Using Video Games to Support Pre-Service Elementary Teachers Learning of Basic Physics Principles. Journal of Science Education and Technology, 20(4), 347-362.
[2a] Bai, H., et al. 2012. Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students. British Journal of Educational Technology, 43(6), 993-1003.
[3a] Coles, C. D., et al. 2007. Games that "work": using computer games to teach alcohol-affected children about fire and street safety. Res Dev Disabil, 28(5), 518-530.
[4a] Din, F. S. and calao, J. 2001. The effects of playing educational video games in kindergarten achievement. . Child Study Journal, 31(2), 95-102.
[5a] Kajamies, A., Vauras, M. and Kinnunen, R. 2010. Instructing Low- Achievers in Mathematical Word Problem Solving. Scandinavian Journal of Educational Research, 54(4), 335-355.
[6a] Kanthan, R. and Senger, J.-L. 2011. The Impact of Specially Designed Digital Games-Based Learning in Undergraduate Pathology and Medical Education. The Impact of Specially Designed Digital Games-Based Learning in Undergraduate Pathology and Medical Education, 135, 135-142.
[7a] Ke, F. 2008. Computer games application within alternative classroom goal structures: cognitive, metacognitive, and affective evaluation. Educational Technology Research and Development, 56(5-6), 539-556.
[8a] Kebritchi, M., Hirumi, A. and Bai, H. 2010. The effects of modern mathematics computer games on mathematics achievement and class motivation. Computers & Education, 55(2), 427-443.
[9a] Ketamo, H. 2003. An Adaptive Geometry Game for Handheld Devices. Educational Technology & Society, 6(1), 83-94.
[10a] Lorant-Royer, S., et al. 2010. Kawashima vs “Super Mario”! Should a game be serious in order to stimulate cognitive aptitudes? Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 60(4), 221-232.
[11a] Miller, D. J. and Robertson, D. P. 2010. Using a games console in the primary classroom: Effects of ‘Brain Training’ programme on computation and self-esteem. British Journal of Educational Technology, 41(2), 242-255.
[12a] Miller, D. J. and Robertson, D. P. 2011. Educational benefits of using game consoles in a primary classroom: A randomised controlled trial. British Journal of Educational Technology, 42(5), 850-864.
[13a] Moreno, J. 2012. Digital Competition Game to Improve Programming Skills. Educational Technology & Society, 15(3), 288-297.
[14a] Moshirnia, A. 2007. The Educational Potential of Modified Video Games. Issues in Informing Science and Information Technology, 4, 511-521.
[15a] Papastergiou, M. 2009. Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation. Computers & Education, 52(1), 1-12.
[16a] Parchman, S. W., et al. 2000. An Evaluation of Three Computer-Based Instructional Strategies in Basic Electricity and Electronics Training. Military Psychology, 12(1), 73-87.
[17a] Poli, D., et al. 2012. Bringing Evolution to a Technological Generation: A Case Study with the Video Game SPORE. The American Biology Teacher, 74(2), 100-103.
[18a] Rastegarpour, H. and Marashi, P. 2012. The effect of card games and computer games on learning of chemistry concepts. Procedia - Social and Behavioral Sciences, 31, 597-601.
[19a] Rosas, R., et al. 2003. Beyond Nintendo: design and assessment of educational video games for first and second grade students. Computers & Education, 40, 71-94.
[20a] St Clair Thompson, H., et al. 2010. Improving children's working memory and classroom performance. Educational Psychology, 30(2), 203-219.
[21a] Suh, S., Kim, S. W. and Kim, N. J. 2010. Effectiveness of MMORPG-based instruction in elementary English education in Korea. Journal of Computer Assisted Learning, 26(5), 370-378.
[22a] Van der Kooy-Hofland, V. A., Bus, A. G. and Roskos, K. 2012. Effects of a brief but intensive remedial computer intervention in a sub-sample of kindergartners with early literacy delays. Read Writ, 25(7), 1479-1497.
[23a] Virvou, M., Katsionis, G. and Manos, K. 2005. Combining Software Games with Education: Evaluation of its Educational Effectiveness. Educational Technology & Society, 8(2), 54-65.
[24a] Yang, Y.-T. C. 2012. Building virtual cities, inspiring intelligent citizens: Digital games for developing students’ problem solving and learning motivation. Computers & Education, 59(2), 365-377.
[25a] Yip, F. W. M. and Kwan, A. C. M. 2006. Online vocabulary games as a tool for teaching and learning English vocabulary. Educational Media International, 43(3), 233-249

CHAPTER 3.
Towards a conceptual framework for assessing the effectiveness of digital game-based learning

Abstract

In recent years, interest has grown in the systematic assessment of the effectiveness of digital game-based learning (DGBL). A conceptual framework describing what effectiveness means in the context of DGBL and which are its subcomponents has hitherto been lacking however. Hence, the goal of this paper is to propose a conceptualization and operationalization of effectiveness rooted in social-cognitive theory. In order to identify desired outcomes and be able to operationalize effectiveness, focus groups were organized with three stakeholder groups following a user requirements analysis methodology.

Results indicate that three categories of desired outcomes can be distinguished: learning, motivational and efficiency outcomes. For the different outcomes, different subcomponents can be extracted which can be organized hierarchically. Learning outcomes that are seen as relevant to the effectiveness of DGBL are 1) increased interest in the subject matter, 2) improvement in objective performance (e.g., in a test), and 3) transfer, referring to the player's ability to apply acquired knowledge or skills to real-world situations. Relevant motivational outcomes concern 1) enjoyment, the extent to which playing the game evoked an enjoyable experience, and 2) increased motivation to learn using DGBL. Efficiency outcomes relevant to DGBL effectiveness, finally, are related to 1) time management and 2) cost-effectiveness. Overall, it can be stated that a DGBL intervention is effective when it achieves similar or higher scores compared to other instructional methods in relation to any of the above mentioned outcomes without significantly (in the common, not the statistical sense) diminishing any of the others.

Keywords:

Evaluation methodologies, interactive learning environments, media in education, interdisciplinary projects

Reference:

All, A., Nunez Castellar, E. P., & Van Looy, J. (2015). Towards a conceptual framework for assessing the effectiveness of digital game-based learning. *Computers & Education*, 88, 29–37.

1. Introduction

Digital game-based learning (DGBL) is increasingly being used in a range of sectors including defense, communication, education, health and corporate training (Backlund & Hendrix, 2013; Michaud, Alvarez, Alvarez, & Djaouti, 2012). Whereas initial research into the topic was exploratory in nature, aiming to demonstrate potential uses, in recent years interest in more systematic assessment of its potential benefits has been growing. An oft-heard argument thereby is that, in order to be considered worthy of investment, research into the effectiveness of digital games as instructional tools is required (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013; Clark, 2007).

Typically, in experimental research on medical treatments, a distinction is made between efficacy and effectiveness (Brook & Lohr, 1991; Flay, et al., 2005; Hunsley, Elliott, & Therrien, 2014). Efficacy refers to the attainment of intended goals under idealized conditions, meaning that experimental control is kept high in order to maximize internal validity (i.e., the interference is an actual result of the treatment and not due to other observed and unobserved variables). Effectiveness also refers to the attainment of intended goals, but under real-world conditions and thus maximizing external validity (i.e., results of an experiment are generalizable to different subjects, settings, experiments and tests), still maintaining an adequate level of internal validity (Brook & Lohr, 1991; Flay, et al., 2005; Hunsley, et al., 2014). Whilst efficacy studies of DGBL interventions are theoretically possible, some flexibility with regard to experimental control will generally be required. Moreover, it is unclear to what extent these studies would provide valuable insights as actual learning generally takes place in less controlled contexts. Hence our primary focus in this paper is on effectiveness rather than efficacy research whilst some observations may also be relevant in the context of highly controlled effect studies.

DGBL generally aims to leverage the entertaining nature of games in order to pursue educational outcomes (Bellotti, et al., 2013). Consequently, with regard to DGBL effectiveness, both learning and player engagement are considered as relevant factors.

1.1. *Defining effectiveness*

Effectiveness of DGBL can be defined as the successful attainment of its intended goals in a real-world context. This implies that its desired outcomes should be clearly defined and made

explicit (Calder, 2013). This is, however, an issue in DGBL effectiveness research as different indicators are being used for determining whether DGBL is effective in different contexts (O'Neil, Wainess, & Baker, 2005). Hence, there is a need for a conceptual model that provides a general evaluation framework for assessment of DGBL which can be applied flexibly across contexts (Mayer, Bekebrede, Warmelink, & Zhou, 2013). In order to move towards a more systematic approach and, consequently, facilitate the comparison of results of different instructional methods across studies, the current study uses social cognitive theory (SCT) as a theoretical framework (Bandura, 1986) to conceptualize and operationalize effectiveness of DGBL. SCT was chosen because it provides a framework of effectiveness evaluation linked to actual behavioral intention as a result of this evaluation. This is in line with previous results stating that proven effectiveness of DGBL will stimulate its implementation (Bardon & Josserand, 2009).

According to social cognitive theory (SCT), motivation for exhibiting a certain behavior is the result of an interaction between personal determinants (cognitive, affective and biological events), behavioral determinants and environmental determinants (Bandura 1986). However, influence of environmental determinants on behavior is not of a direct nature, but is an indirect influence via psychological mechanisms of the self-system. The self-system influences people's aspirations, self-efficacy and personal standards. Or as Bandura states 'What people think, believe and feel affects how they behave' (Bandura, 1986, p. 25).

Bandura's concept of agency provides insight into how effectiveness of human behavior can be evaluated. Agency refers to humans' ability to influence their own behavior through intentionality, forethought and self-regulation by self-reflectiveness and self-reactiveness about their behavior (Bandura 2001). This means that individuals are capable of evaluating their own behavior (i.e., self-evaluation, self-reflectiveness) through observation of that behavior and the associated outcomes (i.e., self-observation). Based on this evaluation, behavior is (dis)continued or altered (i.e., self-reactiveness). This evaluation of behavior occurs based on goal setting which refers to objectives one wished to attain by performing a certain behavior. Hence, outcomes one desired to attain through a particular behavior serves as a benchmark against which to judge effectiveness (Bandura, 2001). If we apply this to digital game-based learning, the evaluation of its effectiveness will be against outcomes one desired to attain by implementing DGBL. Hence desired outcomes of DGBL are considered as the cognitive component influencing behavior, which is implementation of DGBL. Thus desired outcomes of the implementation of DGBL serve as a benchmark for evaluating its effectiveness.

As mentioned before, people's aspirations, self-efficacy and personal standards and consequently, goal setting are indirectly influenced by environmental determinants (Bandura, 1986). Thus, the benchmark against which to judge effectiveness will partly depend on the sector in which it will be implemented and its disciplinary preconceptions (Calder, 2013; Mayer, 2012) which are considered environmental determinants in the present study. Therefore, all relevant stakeholders should be taken into account when aiming at developing an effectiveness definition of DGBL (Calder 2013).

Developing a conceptualization of effectiveness based on desired outcomes of DGBL, indirectly relates it to use. More specifically, if DGBL succeeds in generating outcomes considered relevant for the stakeholders and in validating these outcomes empirically, this will support adoption and usage of DGBL.

Note, however, that not only expected outcomes but also self-efficacy (i.e., perceived capability to perform a certain behavior) influence motivation to perform a certain behavior (Bandura, 1986). Considering that the focus of this paper is effectiveness evaluation and not motivation to use or implement DGBL, (see De Grove, Bourgonjon & Van Looy (2012) and Bourgonjon et al. (2013) for literature on motivations to implement DGBL in a school context), we will focus on outcome expectations in this study.

1.2.*Defining stakeholders*

In order to conceptualize and operationalize DGBL effectiveness, a user requirements analysis with regard to desired outcomes of the implementation of DGBL was conducted. Typically, three types of stakeholders can be distinguished: the operational working area, which refers to stakeholders who will have direct contact with the product, the containing business which refers to stakeholders who benefit from the product in some way, even though they are not in the operational working area and the wider environment, which refers to other stakeholders who have an influence on or an interest in the product (Robertson, 2006). The operational working area was defined as DGBL researchers and game developers, considering that these groups are directly involved in effectiveness assessment whereby the former will conduct the research and the latter prepare the material. The containing business was defined as (potential) adopters of DGBL (e.g., teachers, principals, HR managers, etc.), considering that they would have an interest in the degree of effectiveness of particular interventions they wish to implement. Whilst we considered integrating students, employees, etc., the people who actually play the games, to get an indication on which outcomes they expect from

DGBL, we have decided to put our main focus on intermediaries as adoption of DGBL largely depends on intermediaries such as teachers (Bourgonjon, et al., 2013), considering that the adoption or implementation decision is typically made on a higher level (Boyle, Connolly, & Hainey, 2011; Mayer, et al., 2013).

The wider environment was defined as stakeholders on a governmental level, considering they can have an influence through funding for development of and research on DGBL, of which both DGBL developers and DGBL researchers are mostly dependent on.

2. Method

A user requirements analysis aiming to identify anticipated or desired outcomes of DGBL was conducted. For this purpose, three focus groups were organized; one for each stakeholder group. We conducted focus groups given that it is a cost-effective technique for conducting a user requirements analysis (Maguire, 2003). In total, 33 stakeholders participated in the focus groups (13 in the operational working area, 12 in the containing business and 8 in the wider environment). The participants were recruited by sending out e-mails to eligible people inviting them to participate in the focus groups. We started out with a list of (-location blinded for peer review process-) DGBL developers, DGBL researchers and governmental employees working in education, innovation and media sectors. In our e-mail, we asked for other contacts that belong to one of the three stakeholder groups (for instance, by asking DGBL developers whether they knew companies, schools or health institutions who already have implemented DGBL or are interested in implementing DGBL in the future). Two cinema tickets and a networking lunch after the focus group were provided as an incentive for participating. Table 1 provides an overview of the type and number of stakeholders present in each focus group. During the focus groups, the participants were asked which outcomes they would expect or would like to attain by implementing DGBL and which outcomes would make them decide to implement DGBL. During the focus groups, probe questions were asked to further deepen participants' answers.

Table 1. Overview of participants in focus groups

Stakeholder group	Type of stakeholder	n
Operational working area (N=13)	**DGBL research**	**6**
	- University level	5
	- Corporate	1
	Game development	**5**
	- Location-based gaming company, partly DGBL projects	1
	- Company specialized in games & e-learning development for companies	2
	- Game development company, partly DGBL projects	
	- University for applied sciences for game design	1
		1
	E-learning development	**2**
	- Company specialized in e-learning initiatives for companies, recently started DGBL initiatives	1
	- E-learning platform initiative for primary school	1
Containing business (N=12)	**Education**	**7**
	- Primary school	1
	- High school	2
	- Educational publisher	2
	- Pedagogical guidance	1
	- Support platform culture education for teachers	1
	Corporate	**3**
	- Telecommunications	1
	- Gas transmission company	1
	- Automobile manufacturer	1
	Flemish ministry of mobility and public works	**1**
	- DGBL initiative for road safety	
	Health	**1**
	- Health/sickness fund	
Wider Environment (N=8)	**Flemish ministry for innovation, government investments and media**	**1**
	- Media advisory	
	Flemish ministry for education	**2**
	- Education advisory	1
	- ICT in education policy	1
	Flemish Audiovisual Fund	**1**
	- Game Fund	
	Flemish agency for care and health	**1**
	- Working group for tobacco, alcohol and drugs	
	EU Kids online network	**1**
	- Research	
	Agency geographical information Flanders	**1**
	- Geoservices	
	Flemish agency for culture, youth, sports and media	**1**
	- Gaming policy	

The focus groups were recorded, transcribed and analyzed in qualitative research software package nVivo. The data were analyzed using analytic induction, moving between deduction (i.e., taking into account which elements are currently assessed in DGBL literature and evaluation frameworks in educational research) and induction (i.e., new elements emerging from the data) (Thomas, 2006). The coding took place in three phases; in a first phase the transcriptions were coded at the lowest level, which means segments of texts were labelled

using in vivo coding (e.g., personnel cost, reach, development cost). In a second coding stage, labels referring to similar content were grouped and conceptualized, creating categories of desired outcomes of DGBL (e.g., cost-effectiveness, time management). Finally, three summary categories of desired outcomes were defined: efficiency outcomes; learning outcomes and motivational outcomes. By linking these three core categories together, we developed a multidimensional conceptualization of DGBL effectiveness, for which several indicators can be used in order to operationalize this effectiveness. Table 2 provides an example on how coding occurred based on certain quotes in the data. The labels and categories that were created can be found in the coding scheme in appendix A.

Table 2. Example of how transcriptions were coded (Part 1 of 2)

Summary categories	Categories	Labels	Example of quote
Efficiency outcomes	Time management	Speed of learning	Speed and timing. In the industry it often comes down to teaching people something within a certain timespan. When teaching them something using a computer in comparison to a classical training, one of the goals of the corporation, the client, is to speed up the learning process.
		Achieve more in same timespan	This way [using DGBL] children learn really quickly, even more than I sometimes can achieve in a math course of one hour.
		Reduce time consumption of learners	We also have implemented [non-digital] board games in our company to teach and practice leadership skills. So far, so good, but you also have to make sure that you limit the time consumption of your people, increase the effectiveness and with these board games this was not the case.
		Time required for training teachers	It's important to compare to the usual, more traditional methods. For instance, if you can achieve the same effects a modern for teaching math with a very technical game that aims at teaching math and traditional paper exercises. But if the game costs more and requires more training from the teachers and thus time from the teachers. Then, the investment for the game is much bigger. This can be a way to compare different methods for instance.

Table 2. Example of how transcriptions were coded (Part 2 of 2)

	Cost effectiveness	Cost	I think you also have to take into account certain factors, also budget wise. I think, for instance, that the development of a serious game can cost a lot of money, but when it is developed for the right target group, it can result in a budgetary advantage. A PowerPoint, for instance -which can also be good for instruction- but you have to give this presentation over and over again, a course always has be taught again to new groups and in the end, will cost more money than the development of the game.
		Resources	Someone who is reading out PowerPoints, what is the effectiveness of that? I think that us, companies look at whether there are possibilities to implement more efficient methods regarding resources and personnel.
		Reach	You also have to take a look at how big your reach is, if you only reach 5% of all the schools in a certain region, than maybe it's not worth the cost.

The focus groups were conducted in Dutch. All quotes presented in this paper have been translated as literally as possible and crosschecked by the authors.

3. Results

Our analysis pointed to seven desired outcomes of DGBL. The seven desired outcomes can be subdivided in three categories: learning, motivational and efficiency outcomes. The level of importance attributed to each desired outcome is context-dependent, meaning that the social environment in which DGBL is implemented influences desired outcomes against which to judge effectiveness as we predicted based on our approach rooted in Bandura's social cognitive theory (1986, 2001). For instance, while outcomes related to efficiency and costs are found important in both a corporate and school context, efficiency outcomes related to cost are considered as more important in a corporate context, whereas efficiency outcomes related to time are considered as more important in a school context.

3.1. *Learning outcomes*

3.1.1. Attainment of learning goals as defined by game developer/client

The most crucial element of effectiveness of DGBL is the attainment of the learning goals as defined beforehand by the game developer or the client who ordered the development of the game, which can be teaching (procedural) knowledge, skills or stimulating attitudinal or behavioral change. This, however, implies that a clear description of the learning goal(s) should be provided by the game developers or the clients ordering the development of the game. This description should be specific, meaning that a goal such as 'teaching math' is too vague; 'rehearsing fractions learned in the second grade' provides more information and gives an indication on what should be assessed. Especially the wider environment stressed the importance of providing a measurable formulation of those goals.

> Peter (Wider environment, -blinded- ministry of education, advisor in education): You have to try to make it as clear as possible from the beginning. This means describing the nature of the game, the target group and the intended results. You have to make this as clear as possible to on the whole make it possible to afterwards test whether or not it has led to something and whether or not it is an improvement compared to methods that worked before the game was implemented.

> John (Wider environment, working group for tobacco, alcohol and drugs, policy advisor & funding): When your targets are well-delineated/well-defined, it will be easier to measure your construct. It has to be clear which goals need to be attained and which knowledge should be taught. It would be especially interesting if you look at the attainment of these targets with a game-based approach compared to a handbook or talking to your colleagues.

A possible explanation for why a measurable formulation of the goals is considered important by participants from the wider environment is that in many DGBL projects, particularly those in experimental areas, rely on government funding. Providing a measurable formulation of the goals could make the decision making for providing funding to projects easier, more systematic and more transparent. Moreover, it would create more transparency with regard to valorization, as targets have been formulated beforehand. This was also brought forward by a participant in the operational working area.

> Xavier (Operational working area, university for applied sciences for game design, research coordinator): You have to look at what the aims were and how they were measured. Because that is often the problem: you read something, but you do not know how it was measured.
> Steven (Operatioan working area, serious games development for companies, instructional architect): The aim of companies is to be more profitable; they don't aim at 'peronal development'; it's for instance a matter of how many packages one can produce in one hour. If you could find a way to assess these elements, that would be interesting for these companies.

There was also difference in view on how this should be assessed between the DGBL researchers and the game developers. Firstly, there was some disagreement on the 'scholarliness' of the assessment of learning goals.

> Sophie (Operational working area, experimental psychologist, DGBL researcher): The question is: how can you assess this? Do you look at statistics where you compare test scores between different groups? And when you do this, you should use validated questionnaires or tests, which gives you a more trustworthy study design.

> Oliver (Operational working area, e-learning developer specialized in games & e-learning for companies, manager): I am in the industry, and when we test this, we don't do this in a 'scientific' way, we also don't have this ambition.

Moreover, in the operational working area and the wider environment, preference goes towards comparison of achievement of learning goals with DGBL to a traditional method, such as classroom teaching or paper exercises. However, a participant from the containing business stated that this is rather problematic in a corporate context.

> William (corporate, telecommunications company, training manager): In my company, we would never invest in a training program just to use it as a control group, I have never done this this way.

This is an element that is definitely worth taking into account for further research, as an oft heard critique on effectiveness assessment of DGBL is the lack of a control group where another educational intervention is implemented.

3.1.2. Transfer

Transfer refers to the portability of the learned subject matter in the game to real world settings and situations. This was considered particularly relevant in a corporate and in a health context. The reason for this is that DGBL implemented in these domains mainly aim at teaching skills or changing a certain attitude or behavior. Hence, application of those skills in the actual work environment or a changed attitude towards drugs is the eventual aim and is thus considered more important than test scores.

> Fiona (operational working area, DGBL research, project manager of a serious game developed for a pharmaceutical company): But there's also a difference. You can for instance easily measure if a child gets better in math , but assessing behavioral change is more difficult. A change you might observe in the game -by getting better in the game and thus changing their in game behavior- does not mean they will show a behavioral change in real life.

In a school context this seemed less relevant. This is of course somewhat logical, as the larger parts of subjects treated in a school context are aimed at cognitive learning outcomes or knowledge transfer, of which the aim might not be the application in the real world but more focused on 'general education'. This implies that not only the context in which DGBL is implemented, but also the type of content treated in the game will influence the operationalization of effectiveness.

Transfer is considered as a higher level of effectiveness than the attainment of direct learning objectives, meaning that transfer is considered as a higher order learning effect than objective assessment of performance, by for instance using knowledge tests.

A participant from the containing business referred to the assessment of transfer as a 'strict' way of measuring, whereby one observes in the field whether or not the learner applies what he/she has learned in the game, as opposed to 'lenient' ways of measuring using tests and questionnaires asking for feedback.

> Thomas (Containing business, gas transmission company, training manager): Afterwards, there are two ways of assessing what they have learned. The 'soft way', rehearsing, sending them a questionnaire, actively asking for feedback: is it going better now? Is it working this way? Afterwards, we actually observe: is everything happening like we want it to? A sort of 'inspection'.

Transfer was, however, not brought forward in the wider environment. A reason for this could be that participants of the wider environment relate DGBL to a school context. Also, it could be that assessment of performance using a knowledge test provides enough proof for DGBL effectiveness with regard to learning outcomes.

3.1.3. Increased interest

While attainment of the goals defined by the game developers or the people who ordered the game is considered as an important outcome, it was brought forward in the containing business and the operational working area that a game that succeeds in increasing interest in the subject matter treated in the game would already be considered as an effective game.

> Bart (Operational working area, game developer, owner): For example, in an iPad school in <location removed for blind review process> , in courses where iPads are present and where pupils for instance play games during the math course, pupils will say 'It's math, it's the course with the games'. Maybe they will enjoy it more than another course, where they cannot use the iPad. So, they're much more 'awake' and pay much more attention, because they know they can play a game or a new level of the game…if they are more interested in a subject, then it is already effective for me.

> Evelyn (Wider environment, support platform culture education for teachers, responsible for media literacy): Not only the motivation for playing the game, but also the motivation for the subject matter they have to learn is important: how motivated are they to learn about that subject after playing the game?

Several participants stated that pupils, employees and other learners who start discussing the content treated in the game could be an indicator for increased interest. Another indicator would be the willingness to know more about the subject treated in the game.

William (Containing business, telecommunications company, training manager): But also, if you have played a game, to what extent has the motivation to learn more about the subject increased? I'm thinking that this is something that could be integrated in the post-test. For instance, there has been an article in the newspaper or a documentary on television that covers the same subject covered in the game and you integrate this in your test and you investigate how their interest in knowing more about it has increased, just to give an example. Because in the end, you want a long-lasting effect. We are not going to make hundreds of these serious games. I think if someone in a company has had to play five games, they would say that is enough.

However, stimulating interest is not sufficient for everyone, especially in a more commercial context, where people would have to pay for the game.

Fiona (Operational working area, DGBL research, project manager of a serious game developed for a pharmaceutical company): I think there is a difference between domains for this. In a school context, for instance, showing that the game stimulates interest in the subject might be sufficient because it is one of many resources. In our case, people have to purchase the game, so then you would -especially in this sector- want to see that the game actually works.

Increased interest was also an element that was not brought forward by the wider environment. This gives the impression that for participants belonging to governmental institutions providing funding, performance measures are the most tangible and the most relevant indication for effectiveness.

3.2.*Efficiency outcomes*

3.2.1. Time management

Another element of effectiveness that was brought forward by the operational working area and the containing business is the time span needed for teaching a certain subject matter. If a game helps in reducing the time needed to teach a certain subject matter, resulting in similar learning outcomes, it is considered as effective.

> Oliver (Operational working area, e-learning developer specialized in games & e-learning for companies, manager): Speed and timing. In the industry it often comes down to teaching people something within a certain timespan. If you want to teach them something using a computer in comparison to a classical training, one of the goals of the corporation, the client, is to speed up the learning process.

Time management in a corporate context does not only refer to speeding up the learning process but also to the time required to teaching the whole work force, which is directly related to the cost of the training. Time management is also considered as an important desired outcome of DGBL in a school context. However, perception of this time management is different compared to corporate contexts. While in a corporate context, DGBL is seen as a time reducing replacement for teaching staff with traditional courses, in a school, time management does not only refer to the actual time spent teaching the subject, but also to the achieving more during one lesson hour.

> Geoff (Containing business, Education, ICT teacher): Nowadays, you give a class, the pupils process the content treated in class at home and one week later the teacher does the evaluation. Between the time the class was taught and the evaluation in class, at least two weeks have passed. Whereas with a game, you could for instance see at the end of the class that 20 pupils answered the questions perfectly, so I know I do not have to repeat the subject matter next week. To being able to do all this -teaching the class and evaluating the pupils- as quickly as possible is important for me. Also, the time you have to invest preparing the class is important for teachers. If I would need three weeks to prepare a class using the game and the class only takes 50 minutes, then it is not efficient and it would be better to give the class myself. But if it is something that is ready to use and you can attain several goals with it more quickly, very good, then I do not mind implementing it.

In a school context, also time investment to train the teachers is part of time management.

> Helen (Wider environment, EU kids online network, research): If you can achieve the same effects a modern for teaching math with a very technical game that aims at teaching math and traditional paper exercises, but if the game costs more and requires more training from the teachers and thus time from the teachers, the investment for the

> game is much bigger. This can be a way to compare different instructional methods for instance.

While a decrease in time spent to achieve learning goals is a desired outcome in a corporate context, because it is related to a decrease in cost; in a school context it is desired to reduce preparation and evaluation burdens for teachers. Hence, it is clear that in the corporate context DGBL is seen more as a replacement for traditional classroom teaching, whereas in a school context it is seen as part of the class.

3.2.2. Cost-effectiveness

Cost-effectiveness refers to the cost of teaching a certain number of people a certain subject matter using the game-based method, compared to other instructional methods.

> Cedric (Containing business, car manufacturing company, head of environment department & safety): I think you also have to take into account certain factors, also budget wise. I think, for instance, that the development of a serious game can cost a lot of money, but when it is developed for the right target group, it can result in a budgetary advantage. A PowerPoint, for instance -which can also be good for instruction- but you have to give this presentation over and over again, a course always has be taught again to new groups and in the end, will cost more money than the development of the game.

Whilst indeed, this might be of less significance in a school context, this is still a relevant component of effectiveness assessment in educational contexts, especially in relation to policy. This could, for instance, be an important factor in granting funding for DGBL development projects. When taking into account cost-effectiveness in this decision process, this could also entail that available funding could be accredited to more project proposals.

> Sarah (Wider environment, -blinded- ministry for innovation, media advisor): why should we invest in expensive games when we can get the same learning outcomes with a handbook?

It is clear that cost-efficiency here is still seen in relationship to learning outcomes: at least the same learning effect should be attained compared to the traditional methods. This entails that solely a cost-effectiveness study would not be enough to convince governmental agencies to provide funding for project proposals, but should be accompanied by a comparative study

of learning outcomes, whereas in a corporate context -depending on the company and the content matter treated- reducing costs could be a stand-alone reason for implementing DGBL.

3.3. *Motivational outcomes*

3.3.1. Positive game experience

This element of effectiveness assessment is not directly related to the learning content or the game as an instructional tool, but it is related to the game as an entertainment medium, and more specifically to gameplay and graphical interface of the game. Several participants stated, based on their own experiences with games, that a positive game experience is required to keep pupils motivated to continue playing and consequently, to learn. This was an element that was brought forward in all focus groups.

> Chris (operational working area, company specialized in games & e-learning development, instructional architect): It has to be fun in the sense that it has to be made well, that there are no mistakes in the user interface, because that is just frustrating and then it definitely will not work.

> Evelyn (Containing business, support platform culture education for teachers, responsible for media literacy): To come back to motivation: it is true that you cannot force children to be motivated. A lot depends on how the game looks and whether or not they can link game elements are with their own lives. We once promoted a game and we noticed that it did not work, because it had no links with the lives of the youngsters. After a while, they just stopped playing and we could not motivate them to continue.

According to participants belonging to the containing business and the wider environment, making a fun game does not , however, imply that a very expensive interface should be implemented. Gameplay that is well designed and challenging, is considered to be more important. Whereas according to participants of the containing business -especially developers of commercial games- the criteria used to evaluate commercial games should not be different for DGBL.

Jake (Containing business, education, primary school, principal): I think that recognition and challenging game play makes children motivated to play. It does not have to be very complicated. The best games, such as Tetris, Mario, etc. are the simplest games.

Matt (Wider environment, -blinded- agency for culture, youth, sports and media, gaming policy advisor): I do not know if it is true that games need to be super flashy, because the gameplay is the most important, especially if you are talking about small children. If you sometimes see them working with an iPad, it comes very naturally, a couple of colored boxes are enough so to speak. Thus, regarding development costs it will still be ok.

Jan (Operational working area, game development, manager): Actually, we can make the comparison with commercial games and how these are evaluated, which is based on reviews which in turn are based on certain criteria: is the game fun? Are the graphics nice? Is gameplay challenging? I think we can definitely transfer certain of these criteria to serious games, but serious games will have an extra criteria: does it achieve its goals? Was it instructive? I think we would want to work towards something like that and I think that is possible.

Thus, all stakeholder groups in the present study are following the idea of DGBL's twofold goal: learning and entertainment and find it important that a game is effective in both respects. However, how this entertainment is perceived differs between stakeholder groups. Participants from the containing business and wider environment seem to perceive game experience as secondary to learning; the gameplay should thus be engaging enough to support learning and keep learners motivated to play, while game developers see it as a priority to create a positive game experience. This is, of course, logical considering that creating entertaining and graphically nice games are a main part of game developers' expertise and a way of bringing forward this expertise. Moreover, a primary focus on creating a positive game experience is especially relevant for commercial of the shelf serious games.

Bart: (Operational working area, game developer, owner) It just has to be fun for the target group and hopefully, they will learn something. But if you are focusing too much on what they have to learn and achieve, then it won't work, I think…This is especially important for a commercial game, so they keep playing or want to replay to get higher scores for instance.

A main focus on fun is, however, considered a major issue on governmental level; stating that DGBL sometimes lacks a theoretical base for instructional techniques, which already should be thought about during development.

> Ruth (Wider environment, -location removed- Adiovisual Fund, project management animation & games): Research before you start developing the game, that is also very important.

> Josh (Wider environment, (-blinded- agency for care and health, working group for tobacco, alcohol and drugs): The problem is that a lot of people start from a creative perspective… An analysis should be made of which instructional methods work for a certain problem.

3.3.2. Motivation towards the instructional method

Motivation to learn using the game-based method is a desired outcome of DGBL brought forward by the containing business and the wider environment. Moreover, some participants stated that when motivation for learning through DGBL is higher compared to a traditional instructional method this would be a decisive factor for implementing DGBL. However, a prerequisite for this would be that similar levels of learning outcomes are attained. Motivation towards the instructional method is thus an important component to take into account when assessing effectiveness of DGBL, considering that it plays a role in the implementation decision of DGBL for potential adopters and it could play an important role in funding decisions.

> Sarah (Wider environment, -blinded- ministry for innovation, media advisor): If you can say that the game improves knowledge and that the motivation to learn using the game is higher compared to the handbook, than it has a clear added value compared to the other tool.

> Jasmine (Containing business, education, educational publisher): The reason why we developed our math game a couple of years ago, is because there was a larger need from the field [eduation]. The ‘drill’ isn’t there anymore and children aren’t motivated anymore to do paper drill exercises. That is why we started looking for other ways and came up with a game.

What is surprising, is that motivation towards the instructional method was not brought forward in the operational working area. Hence, in the present study, priorities of DGBL developers and researchers are not entirely in line with those of potential adopters or policy, considering that the participants themselves indicated that this would be an important factor in deciding to implement DGBL. This could be due to the fact that DGBL developers and researchers are rarely directly involved in the implementation of the actual game. Consequently, these stakeholders might be less sensitive to challenges adopters are facing such as motivating personnel to engage in supplementary training or a refresher course.

4. Discussion

Our study shows that conceptualizing DGBL effectiveness is possible using an approach based on desired outcomes. By using an approach rooted in socio-cognitive theory, an evaluation framework for DGBL effectiveness linked to actual implementation of DGBL was provided. By linking our results to literature we have developed an operationalization of DGBL effectiveness which can be found in table 3. This is not only relevant for DGBL researchers, providing them with relevant outcomes to assess, but it can also be useful for DGBL developers, providing them with desired outcomes of relevant stakeholder groups and thus elements to take into account when developing DGBL in order to stimulate its implementation. Moreover, by conducting a user requirements analysis involving different stakeholder groups, our study has shown that the importance that is attributed to desired outcomes depends on the sector in which it is implemented.

Our results bring forward a multidimensional conceptualization of DGBL effectiveness, suggesting the existence of three categories of outcomes that should be considered when assessing DGBL effectiveness: learning, motivational and efficiency outcomes. For these outcomes, several indicators can be used. There is, however, a hierarchical relationship between these indicators.

The most important element that was brought forward with regard to effectiveness of learning outcomes by all stakeholder groups was the attainment of learning goals embedded in the game as defined by the game developer or client who ordered the development of the game. It was, however, brought forward that a game that succeeds in stimulating interest in the subject treated in the game, but cannot objectively demonstrate an improvement in performance, could already be considered as an effective game. Therefore, an increased

interest can be seen as a first level of effectiveness with regard to learning outcomes. In the academic literature, this is defined as situational interest. In contrast to individual interest, which has a dispositional quality and, consequently, is relatively constant in different situations, situational interest is a result of responses to features of the environment in which the content is taught (Linnenbrink-Garcia, et al., 2010). Situational interest consists of two elements: triggered situational interest, which is an attentional reaction to the environment and aims at grabbing the learner's attention and maintained situational interest, which is an affective reaction to the environment, which keeps the learner interested as a result of a 'deeper' connection with the content and an increased perception of its significance to the individual (Linnenbrink-Garcia, et al., 2010). The second level of effectiveness with regard to learning outcomes is the objective assessment of performance after having received the game-based intervention, looking at whether or not learning goals defined by the game developers (or the client who ordered the game) are achieved. A third level of effectiveness with regard to learning outcomes is transfer, which is especially relevant in game-based interventions aiming at teaching or training a certain skill or behavior (e.g., a game aimed at drug prevention). Transfer is, however, an element that is currently not regularly being assessed in DGBL effectiveness studies (Mayer, et al., 2013). This thus merits more attention in the future. Key performance indicators, which refer to the gathering of data in the field (for instance, the number of accidents on a monthly basis when doing digging works) could be a possible way of assessing transfer (Mayer, et al., 2013).

Effectiveness with regard to motivational outcomes consists of motivation towards the instructional method and a positive game experience. While game-experience is not directly related to the game as an instructional tool, a positive game-experience stimulates positive reactions towards the media and its contents (Vorderer, Klimmt, & Ritterfeld, 2004) and can influence motivation towards learning via the game-based approach, resulting in increased progress (Giannakos, 2013). Hence, game experience can be considered as the first level of motivational outcomes. Game experience is defined as 'an ensemble made up of the player's sensations, thoughts, feelings, actions and meaning-making in a gameplay setting' (Ermi & Mäyrä p.2) and is thus a complex concept. Moreover, considering that it points to an interaction between game, gamer and context, it is agent dependent and thus subjective (De Grove, Bourgonjon & Van Looy, 2012). Consequently, it is suggested that it is better to look at game experience as an underlying mechanism that make games motivating and fun, which stimulates performance. Therefore, in order to get an indication of game experience, our results revealed that a general suggestion would be to focus on the concept of enjoyment (De

Grove, Bourgonjon & Van Looy, 2012), which refers to a positive game experience, a feeling of being entertained or in short, a game that is perceived as 'fun' (Moser, Fuchsberger, & Tscheligi, 2012). However, accurate instruments to assess several aspects of the game experience could give game developers an indication on where the game fails to create a positive game experience.

Effectiveness with regard to efficiency outcomes relates to making good use of resources available (Calder, 2013) and is thus related to outcomes on the organizational level, more specifically it is related to return on investment of the game-based program. In our results, this is associated with cost-effectiveness and time management. While cost-effectiveness was especially relevant in a corporate context, time management was relevant in both a corporate and educational context. This is an element, that is rarely brought forward in DGBL effectiveness studies while it is, however, a suggestion that has already been brought forward in a criticism by Clark (2007), considering that DGBL effectiveness studies often find no significant difference if compared to another treatment. The added value of the DGBL treatment should then be reflected in the difference in cost of the treatment or the time gained by the treatment, which again, can indirectly be brought back to a decreased cost. Our results also confirmed that a decreased cost and/or time play an important role in the decision for implementation of DGBL.

Table 3. Operationalization of DGBL effectiveness

Learning outcomes	Motivational outcomes	Efficiency outcomes
Situational interest **DGBL stimulates interest in the content matter discussed in the game.**	***Enjoyment*** DGBL succeeds in creating an enjoyable game experience.	***Time management*** DGBL succeeds in reducing the timeframe required to teach a certain content matter. This is a judgment of relative worth compared to other instructional methods.
Performance **DGBL succeeds achieving learning goals as defined by the game developer/the client who ordered the development of the game.**	***Motivation towards DGBL*** Learning with the game-based method is motivating. This is a judgment of relative worth, compared to other instructional methods.	***Cost-effectiveness*** DGBL succeeds in reducing the cost of the intervention with regard to: a) the number of learners that can be reached and b) the time required to teach the target group certain content. This is a judgment of relative worth, compared to other instructional methods.
Transfer **DGBL stimulates application of learned content matter in the game to real world situations.**		

Overall, it can be stated that a DGBL intervention is effective when it achieves similar or higher scores in comparison to 'business as usual' in relation to any of the above mentioned outcomes without significantly (in the common, not the statistical sense) diminishing any of the others. This does not imply that all outcomes need to be assessed. This depends on the sector in which it is implemented and the type of content treated in the game. Consequently, one does not necessarily need to assess situational interest in order to assess performance. For instance, a company might be interested in replacing their current fire safety class by DGBL. Reasons for this are primarily decreasing costs but not at the expense of performance. Moreover, by providing a game that resembles the company environment, they aim at portability from what they have learned in the game to the real word. Lastly, they hope the game will be a more motivating tool to learn about fire safety than the current class. Consequently, the effectiveness evaluation for the game the company ordered will need to focus on performance, motivation towards DGBL, transfer and cost-effectiveness. Hence, table 2 can be used as a communication tool between DGBL developers, DGBL researchers and clients in order to tune game development and research to the needs of the client ordering the development of DGBL. By doing this, implementation of DGBL is stimulated considering that according to social cognitive theory more outcomes are desired for the client(s), the higher the likelihood of implementing DGBL (Zimmerman & Schunk, 2003).

5. Limitations and further research

A limitation of the present study is that our sample is not representative for the populations of the different stakeholders. While operationalization of DGBL effectiveness will always be dependent on the specific use contexts, the present study did succeed in identifying certain general components. An interesting venue for further research would be an extension of the operationalization of effectiveness based on desired outcomes of other potential adopters, such as parents and individuals who decide to learn a certain content matter using DGBL. Also, a further validation of this framework based on quantitative methods such as survey research would be useful. While assessment of the outcomes was outside the scope of this research, this would be a next step in creating a more systematic approach for the assessment of DGBL effectiveness.

CHAPTER 4.
Assessing the effectiveness of digital game-based learning: best practices

Abstract

In recent years, research into the effectiveness of digital game-based learning (DGBL) has increased. However, a large heterogeneity in methods for assessing the effectiveness of DGBL exist, leading to questions regarding reliability and validity of certain methods. This has resulted in the need for a scientific basis to conduct this type of research, providing procedures, frameworks and methods that can be validated. The present study is part of a larger systematic process towards the development of a standardized procedure for conducting DGBL effectiveness studies. In a first phase, the variety in methods that are used for sampling, implementation of the interventions, measures and data analysis were mapped in a systematic literature review using Cochrane guidelines. The present paper reflects the second stage, where this variety in elements are presented to experts in psychology and pedagogy by means of semi-structured interviews, in order to define preferred methods for conducting DGBL effectiveness studies. The interview was structured according to five dimensions that were used in the literature review: 1) participants (e.g., characteristics of the sample involved) 2) intervention (e.g., contents, format, timings and treatment lengths, intervention(s) in control group(s)) 3) methods (sampling, assignment of participants to conditions, number of testing moments) 4) outcome measures (e.g., instruments used to measure a certain outcome) and 5) data-analysis.

The interviews were transcribed and analyzed using qualitative software package nVivo. Our results show that areas for improvement involve the intervention dimension and the methods dimension. The proposed improvements relate to implementation of the interventions in both the experimental and control group, determining which elements are preferably omitted during the intervention (such as guidance by the instructor, extra elements that consist of substantive information) and which elements would be aloud (e.g., procedural help, training session). Also, variables on which similarity between experimental and control condition should be attained were determined (e.g., time exposed to intervention, instructor, day of the week). With regard to the methods dimension, proposed improvements relate to assignment of participants to conditions (e.g., variables to take into account when using blocked randomized design), general design (e.g. necessity of a pre-test and control group) test development (e.g., develop and pilot parallel tests) and testing moments (e.g., follow up after minimum 2 weeks). In sum, the present paper provides best practices that cover all aspects of the study design and consist of game specific elements.

While several suggestions have been previously made regarding research design of DGBL effectiveness studies, these often do not cover all aspects of the research design. Hence, the results of this study can be seen as a base for a more systematic approach, which can be validated in the future in order to develop a standardized procedure for assessing the effectiveness of DGBL that can be applied flexibly across different contexts.

Keywords:

Digital game-based learning; Evaluation methodologies; Evaluation of CAL systems; Interactive learning environments; Media in education

Reference:

All, A., Nunez Castellar, E. P., & Van Looy, J. (2016). Assessing the effectiveness of digital game-based learning: best practices. *Computers & Education, 92-93,* 90–103.

1. Introduction

Digital games encompass a variety of types and genres that can be played using a multitude of digital technologies such as computers, (handheld) consoles and mobile devices. Based on a literature review on digital games definitions, Juul (2003) defines a digital game as

> …a rule-based formal system with a variable and quantifiable outcome, where different outcomes are assigned different values, the player exerts effort in order to influence the outcome, the player feels attached to the outcome, and the consequences of the activity are optional and negotiable (p.5).

Digital game-based learning (DGBL) refers to the usage of the entertaining power of digital games to serve an educational purpose (Prensky 2001). DGBL is the consequence of a balance between learning and gaming elements (Nussbaum and Beserra 2014). DGBL contains two important elements: fun/entertainment and an educational component (Bellotti et al. 2013). Consequently, in the DGBL literature and published effectiveness studies both learning and player engagement/motivation are considered relevant to assess (Bellotti et al. 2013).

Two types of games can be distinguished in DGBL: special purpose games which have been developed with an educational purpose and Commercial-Off-The-Shelf games that have been developed for entertainment purposes, but that are being deployed in an educational context. Note, however, that this does not mean that special-purpose DGBL games cannot be commercially available (Stewart, 2013)

Based on the projected primary learning outcomes, three types of special-purpose games can be distinguished. They aim to achieve knowledge transfer (cognitive learning outcomes), skill acquisition (skill-based learning outcomes), and/or attitudinal/ behavioral change (affective learning outcomes) (Stewart, et al., 2013). Games that are developed with the primary aim of achieving knowledge transfer are typically implemented in education, in order to teach math (Castellar, All, de Marez, & Van Looy, 2015) or language (Yip & Kwan, 2006), for example. Digital games that primarily aim to support skill acquisition are generally used for training, for example in a corporate or military context. For instance, several studies have examined the impact of playing games to develop managerial skills (Corsi, et al., 2006; Kretschmann, 2012). Games that are developed to achieve attitudinal change are sometimes

used by governments and NGOs to raise awareness about a certain topic, such as poverty (Neys, Van Looy, De grove, & Jansz, 2012). Games with a behavioral change intention are typically found in the health sector. For example, some games promote healthy food and physical activity to children (Baranowski, Buday, Thompson, & Baranowski, 2008). While DGBL can primarily aim to achieve a certain type of learning outcome, this does not exclude secondary learning outcomes (Kraiger, Ford, & Salas, 1993). For instance, a game that primarily aims to teach children English (cognitive learning outcomes) can also result in a more positive attitude towards learning English or English as a subject (affective learning outcomes).

Although meta-analyses have proven the effectiveness of DGBL (Backlund & Hendrix, 2013; Clark, Tanner-Smith, & Killingsworth, 2015; Connolly, Boyle, MacArthur, Hainey, & Boyle, 2012), certain authors have pointed out elements that jeopardize reliability and validity of some findings. This includes comparisons with control groups that did not receive an educational intervention (Hays, 2005), time-on-task differences between experimental and control groups, and validity of research instruments (Randel, Morris, Wetzel, & Whitehill, 1992). Moreover, some studies do not provide enough information about the implementation of the intervention (Clark, 2015; Sitzmann, 2011). This makes it hard for readers to know if the reported results are a consequence of the different methods, and not a cause of circumstantial factors that differed between conditions (Randel, et al., 1992). Rigorous assessment is required to improve the quality of DGBL, to support resource allocation, and to gain insight in the most effective way to use games to support learning (De Freitas, 2006, Kirriemur, 2004).

1.1. *Studies about DGBL effectiveness*

Two types of evaluation of educational interventions can be distinguished. A first type is formative evaluation which aims to determine areas for improvement and is thus an evaluation of the process of the intervention itself. This type of evaluation is conducted by using a naturalistic design with observational data collection, which describes an ongoing process in its natural setting. A second type is summative evaluation, which aims at to determine whether or not an educational intervention succeeds in attaining its goals, thus evaluating the outcomes (Calder, 2013). Summative evaluations are conducted by using an experimental design (Hutchinson, 1999). In the present study, we focus on summative evaluation and will concordantly discuss experimental design.

An earlier content analysis on the effectiveness of DGBL approaches, conducted by the current authors, showed that there is a large diversity in the way that experimental research on DGBL effectiveness assessment is conducted, making comparison of results across studies difficult. This heterogeneity can be found on all four dimensions of the study design, as defined by Cochrane guidelines, which were used for the content analysis (i.e., a systematic review method which has its origins in health research and aims to assess the effectiveness of interventions for prevention, treatment and rehabilitation (Higgins 2008). The dimensions are 1) participants (e.g., characteristics of the sample involved), 2) intervention (e.g., contents, format, timings and treatment lengths, intervention(s) in control group(s)), 3) methods (e.g., applied research methods) and 4) outcome measures (e.g., instruments used to measure a certain outcomes). Variety is caused by three main issues: the type of activity implemented in the control group (no activity, traditional classroom teaching, computer-based learning, other game, etc.), the outcome measures that are used to assess effectiveness (perceived learning, time on task, test scores, student achievement, etc.), and different statistical techniques that are used to quantify learning outcomes (percentage of improvement, between group comparison with repeated measures, post-test scores comparison, etc.) (All, Nuñez & Van Looy, 2014). Table 1 provides a more detailed overview of the main differences between studies on DGBL effectiveness.

Table 1. Main differences across DGBL effectiveness studies regarding methodology.

Aspect of study design	Main differences across studies (N=25)
Participants	• Large variety in sample size • Reporting on types of people included
Intervention	• Activity implemented in control group(s) • Stand-alone intervention vs. embedment in a larger program • Variety of elements present in larger program • Presence of / role of / type of intermediary
Method	• Assignment of subjects to conditions • Use of matching/ blocked randomized design in different ways • Addition of qualitative data
Measures	• Different objective measures of performance • Different self-report measures • Similarity pre- and post-test(s) • Data-analysis techniques

Adapted from All, Nuñez Castellar & Van Looy (2014).

Results of the content analysis also revealed certain suboptimal study designs which are related to confounding elements. Three main issues can be distinguished. Firstly, the addition of elements to the game, such as required reading, extra exercises, or debriefing sessions, makes it impossible to isolate the effect of the game. Secondly, the type of instructor present during the intervention (familiar vs. unfamiliar person) and the role the instructor has during the

intervention differs across studies. Instructors are either present to 1) only supervise, 2) offer technology oriented support when respondents encounter issues concerning the technology or actual game play (i.e., procedural help), or 3) offer content-related help, by providing contextualization of game play and in game elements in the broader learning context during actual game play (i.e., guidance) (All, Nuñez Castellar & Van Looy, 2014). Thirdly, implementation of the same test pre- and post-intervention on the same day, could lead to practice effects and pre-test sensitization. This would, again, result in an overestimation of the instructional effect (Van Engelenburg, 1999, Crawford, 1989). In 1992, Randel mentioned similar issues with regards to the reliability and validity of certain effectiveness studies on instructional games. Twelve years later, the same issues are still detected in DGBL effectiveness research.

1.2.*Towards an overarching methodology*

The heterogeneity in study designs, which leads to mixed results and critiques on certain study characteristics, has resulted in a research field which is unable to make generalized claims about the successfulness of DGBL (Giessen 2015). An underlying reason for this is that DGBL is an emerging field, which combines different disciplines with specific research traditions (Kirriemuir & McFarlane, 2004; Mayer, et al., 2014). More specifically, evaluation of DGBL effectiveness is at the crossroads of psychology and pedagogy (Connolly 2014). Hence, there is a need for an overarching methodology to research and evaluate DGBL, which should provide procedures, frameworks, and methods that can be validated (Mayer et al. 2014). While several suggestions have been made to improve the design of DGBL effectiveness studies (Mayer et al. 2014, Serrano-Laguna et al. 2013), these do not cover all aspects of the experimental research design (e.g., aspects for which similarity between subjects should be attained, instructor role, etc.).

A common methodology would firstly create the opportunity to compare results, and thus the quality, of the different instructional methods across studies. Secondly, claims regarding the effectiveness of DGBL could be made on a more generalized level: per field (e.g., science, math, language learning) or per game genre. Thirdly, a common methodology would set a baseline for quality, which could serve as an evaluation tool for published studies and as a starting point for researchers wanting to conduct a DGBL effectiveness study. Lastly, interest in studying game design features (e.g., competition, narrative, etc.) that influence effectiveness, is growing in order to optimize DGBL game design. In order to make general

claims about game design features that influence effectiveness, a standardized approach for studying effectiveness is required (Cagiltay et al. 2015, Kirriemuir & McFarlane, 2004).

As the field of DGBL effectiveness research is relatively new and no guidelines currently exist for conducting these types of studies, expert interviews are considered appropriate for defining best practices. Expert interviews are typically used for exploration in an emerging field and allow the accumulation of both process and context knowledge (Flick, 2009). Experts possess both theoretical knowledge and experience with actually executing experimental research. Thus, they know how to tackle some issues related to experimental research in a DGBL context. The present paper aims to formulate best practices based on expertise coming from both experimental research in both psychology and pedagogy in order to create a more standardized evaluation approach.

In the present study, as part of a larger process towards a more standardized approach for conducting DGBL effectiveness studies, we focus specifically on special purpose DGBL. Considering the different types of learning outcomes require different types of assessment (Kraiger, et al., 1993) and thus resist categorization in one research taxonomy, we solely focus on cognitive learning outcomes.

2. Method

Experts have been defined as *'staff members of an organization with a specific professional function and a specific experience and knowledge for this purpose'* (Flick 2009 p.166). Based on this description we selected professionals with at least a Ph.D. degree in either educational sciences or psychology who have conducted, or are still conducting, relevant research which evaluates educational interventions. We used a combination of purposive sampling, based on the criteria stated above, and snowballing (Flick 2011). Interviewed experts were requested to provide other experts who could be relevant for this study. Seven experts in psychology (five national and two international experts) and six experts in pedagogy (two national and four international experts) were interviewed. Ten interviews were conducted using videoconferencing software and three were conducted face-to-face.

Interview questions were derived from a review study conducted by the authors (see paragraph 1.1.). The interview was structured according to four dimensions of Cochrane guidelines that were used in the literature review: 1) participants 2) interventions 3) methods

and 4) outcome measures. We have added a fifth dimension, data-analysis, as the content analysis of the authors indicated that different statistical techniques are used to quantify learning outcomes in DGBL effectiveness studies. Hence, the interview was conducted according to five instead of four dimensions. The interview guide can be found in Appendix A.

The interviews were conducted over a period of two months. During this period, interviews were transcribed and analyzed using the qualitative analysis software package nVivo. The transcribing and coding did not occur at the end of the process. Instead, the 'constant comparison' principle was applied: this refers to simultaneous relationship between collection and analysis of data (i.e., not a sequential relationship where all interviews are conducted first followed by analysis of all these interviews) (Suddaby, 2006; Glaser & Strauss, 2009). This allowed us to conduct interviews until 'category saturation' was achieved (Strauss & Corbin, 1990). Signals of saturation are the '*repetition of information and confirmation of existing categories'* (Suddaby, 2006, p. 639*).* Hence, data collection stopped when no new codes were developed during the analysis. Thirteen semi-structured interviews were conducted in total, which is considered an acceptable sample size for expert interviews (Baker and Edwards 2012).

The interviews were analyzed inductively and the experts' answers and comments were coded in three phases (Thomas 2006). In a first phase, the transcriptions were coded at the lowest level. This means segments of texts were labeled using in vivo coding (i.e., defining the text by a concept/phrase used by the interviewee). In a second coding phase, labels referring to similar content were grouped and conceptualized, creating categories (e.g., similarity between conditions). In the last phase, these categories were attributed to dimensions of the study design (e.g., research design, participants, intervention, outcome measures and data analysis). An example of how coding occurred can be found in table 2.

Table 2. Example of how transcriptions were coded.

Phase 3 (Dimensions of the study design)	Phase 2 (creating categories)	Phase 1 (in vivo coding)	Example of quote
Intervention	Similarity between experimental and control group	Equal time exposed to intervention	All these interventions should however be matched on different criteria, for instance: content, time of exposure
		Same content in conditions	If you have more difficult questions on the paper and pen exercises; then yeah… the content should in principle be the same. The subject of study is a medium, so the content in both should be identical.
		Same interaction with other people across conditions	The amount of interaction one has with other people, I do not know if this is always matched in both conditions. That seems relevant to me.
		Support received	If you think as an instructor that this intervention requires explanation, above what you would usually give if you use the conventional method, this might trigger better learning. But had you given the same explanation to the conventional method, the effect would have been the same, right? So, there have to be rules on how much you reveal or how much you help people because it's a new environment and this level of help and support and assistance has to be comparable across this two treatments. And I agree that the game methods, the educational game treatment will require maybe some more help. But it has to be specified very clearly why you use help, why you intervene, and if you do, it would be important to do it for everybody. Right?
		Same instructor across conditions	… whether it was the same instructor who instructed the people in the game group and in the control group because it might be an instructor effect if these are different people, right?

3. Results

3.1. *Research design*

3.1.1. Control group

Experts were asked what the most ideal design would be, when assessing the effectiveness of a DGBL intervention aimed to achieve cognitive learning outcomes. The majority (11/13) of the experts expressed the need for a control group. This would allow evaluation of two aspects: (1) if positive outcomes are related to the mere lapse of time and (2) the comparison of motivational outcomes.

Two experts, however, stated that there is no need for a control group. According to them, determining whether or not predefined goals are attained would be sufficient. The reason for this is that firstly, a control group is only required when the research question involves a comparison with another method. Secondly, many differences can exist between the educational intervention the game is compared to and the game itself, resulting in a flawed comparison.

It was, however, generally agreed that when a control group is included in the research design, another educational activity should be implemented in the control group. Comparing with a control group that does not receive instruction only results in knowing whether being exposed to certain content through DGBL games is better than not being exposed to any learning content at all, which does not add value to either the research field or society. It would be most preferable to compare with 'business as usual'. In the context of educational research this could, for instance, be traditional classroom teaching, doing exercises, or the use of another electronic platform (if this is how pupils are currently being instructed, for example in long-distance education). The implementation of another digital game in the control group, not related to the learning content, would only be interesting when examining the motivational rather than the cognitive aspects of DGBL. One expert stated that another game could also be implemented as a third condition.

> That way, you can examine whether or not the combination game and math is of added value. If you for example find higher results for condition 1 (combined condition game + learning content), compared to condition 2 (learning content in a traditional way) and condition 3 (game without learning content), you know the combination of the two elements is superior. If you only find a difference between condition 1 and condition 2, it can be attributed to the fact that it was a game, and therefore participants were more motivated. And if you only look at the difference between condition 1 and condition 3, you could say that it is due to the integrated learning content in the game. While if you find this result, you know the combination has an added value. This is a conclusion you cannot reach while only working with condition 2, or condition 3 compared to condition 1.
>
> (Experimental psychologist, Professor in experimental methods, Belgium)

One expert stated that the comparison with another, non-digital game would be of interest, considering these entail similar processes.

> With hindsight, I now favor the idea of having - and one of my doctoral students has just done a study using this approach - the controls having a parallel, if you like, or a

> similar learning experience which addresses the same learning outcomes and the same processes, but not using technology. In fact, she [doctoral student] had the children using card games. It is just another interesting way of looking at it: any differences can then be attributed not to extra learning or the introduction of new concepts, but to the actual technology.
> (Educational scientist, Professor in educational studies and research methods, Scotland)

3.1.2. Pre-test

A pre-test was generally considered indispensable in this type of research for three reasons. Firstly, when no pre-test has been implemented, differences between the experimental and control group at the beginning of the intervention are not controlled. This means that the experimental group could, by chance, have had higher levels of knowledge before the intervention, resulting in an overestimation of the effect of the game-based learning intervention. Secondly, a pre-test is necessary in order to determine the relative learning gains of the participants as a result of the intervention. Thirdly, when no pre-test is implemented, there is no control for characteristics of drop-outs, which is especially relevant in the context of educational research.

> What you often see in this type of studies, is that drop-out is not random; especially participants who performed poorly, don't feel like participating anymore. They might find it confrontational to come back, because they feel ashamed.
>
> (Experimental Psychologist, Professor in data-analysis and statistics, Belgium)

Two experts also stated that the ideal design would be a Solomon 4-group design, creating four conditions, of which two conditions -one in the experimental and one in the control group- do not receive a pre-test, in order to control for practice effects (i.e., when taking the same test twice, participants generally do better the second time). One expert did, however, state that a pre-test is not always necessary, if the learning goals of the intervention are clearly defined. When no pre-test is administered, randomization of subjects should be applied.

3.1.3. Follow-up study

Furthermore, the vast majority (11/13) of the experts considered the integration of a follow-up study as good practice. This is especially relevant with short interventions, in order to examine whether or not the effect was a result of intensive training. Several experts indicated that in educational research, effects that disappear after a few days have little use. A minimum 'longer term' assessment required would be two weeks. Ideally, the period would be three to six months and up to one year for longer interventions. Moreover, instead of announcing the follow-up study, a surprise recall would increase ecological validity. The experts, however, are aware of the fact that organizing a follow-up study after, for instance, one year is rather difficult in practice due to attrition (i.e., loss of cases over time).

3.1.4. Assignment of participants to conditions

Randomization has been accepted by all experts as the preferred method in order to keep groups as similar as possible in terms of gender, age, motivation, etc. Two types of randomization were discussed: randomization of subjects and randomization of classrooms. While experts mention randomization of subjects as the most preferable method, they acknowledge that this is not always possible in real life for two reasons. Firstly, randomization of subjects entails the need for a larger sample size, which is often an issue in this type of research. Secondly, randomization of subjects is not always possible due to the context of the study (e.g., implementation in a natural collective such as a class group). Due to the practical limitations mentioned above, randomization at the classroom level would be another option. A limitation of this type of randomization is that school influences could result in a biased sample.

> The main problem with this [randomization on the classroom level] is that you might bias your sample; if for instance you have a school where only pupils of low socio-economic status go, than if all the classrooms come from this school, results might not be biased. You just have to make sure that you mention that these results are only valid for children with low SES. If you have mixed classrooms, which mix children with a high SES and a low SES, and by chance you only give the manipulation to children with high SES... If you conclude that the manipulation worked, you might have biased results.
>
> (Experimental psychologist, Researcher on digital-game based learning, Belgium)

Furthermore, classroom influences related to teacher characteristics (e.g., teacher style, experience with and attitude towards games) could lead to biased results. Matching has also been suggested by the majority of experts (12/13) as a way of guaranteeing similarity between conditions, controlling for certain variables.

> I think matching on characteristics is better than doing no matching, but it's not as good as randomization. There could be unobserved characteristics that you are unable to match on. So, it could be a matter of student motivation. It could be that only the most motivated and hardworking students are going to sign up for an online course versus a traditional course because it requires more self-motivation on the part of the student to get it done without a teacher there watching them every minute. So you do have to worry about differences in unobserved characteristics if you match on characteristics like test scores. As I mentioned earlier, I have done some other studies that have used matching and I think it's better than not doing any type of control for comparison conditions, but it's not as good as a randomized controlled trial. (Educational Scientist, Senior research scientist, assessment of educational interventions, US)

Table 3 provides an overview of variables to match participants in different conditions as suggested by the experts.

Table 3. Variables suggested to match on

Variable	Description
Previous knowledge	Matching on prior academic achievement or pre-test scores
Ability	Matching on different ability levels (e.g. low, medium and high achievers)
Motivation	Matching on motivation towards the learning content
Game experience	Matching on previous experience with games
Gender	Matching on gender (male/female)
Age	Matching on age/age categories
SES	Matching on socio-economic status

3.2.*Participants*

3.2.1. Sample Size

An absolute minimum suggested by the experts is 20 participants per condition. For more sophisticated statistical analyses, a required minimum would be 30 participants per condition. Several experts suggested conducting a power calculation beforehand. This would serve as a basis for determining how large the sample size should be in order to detect a real difference (and not miss them). Assumptions would have to be made about the magnitude of the effect (e.g., effect size) to determine power. The calculation itself could be added in an appendix.

3.3. *Intervention*

3.3.1. Context of the intervention

With regard to context of play, expert opinion is divided. Four experts have a strong preference for implementing DGBL in a formal context. This creates opportunities for more control and should result in a higher internal validity. The other experts have a preference for a context that is representative for the game implemented, in order to increase ecological validity.

> 'I think that you should describe, as much as possible, how you implemented the game. At least there should be some reference to a website, for instance, where the game and intended gameplay is described, in case there is not enough room in your article.'
>
> (Educational scientist, Professor in research methods and statistics, The Netherlands)

With regards to an informal gameplay context (e.g. at home), certain issues about control are raised. Nevertheless, control should be possible in this context according to one expert.

> 'I find it no problem that children play a game at home, but you should be sure that they do it under certain conditions, such as under parental supervision, and that you can also define that they play during specific hours or a maximum during a time period for instance.'
>
> (Cognitive psychologist, researcher on digital-game based learning, Belgium)

Logging could also be an opportunity, but by itself it does not ensure control.

> 'You can log what they do and can keep track of how long they are playing…but still, then you don't really know if they just leave the game 'open' and go for a drink in the meantime for instance. So you still have some things you can't control.'
>
> (Educational scientist, Professor in educational technologies, Belgium)

There is a general agreement however, on conducting these types of studies in a lab environment. The lab is a very controlled way of doing research, which will not be representative for the real world. A lab study could, however, serve as a first phase in research.

3.3.2. Similarity intervention and control group(s)

One of the major limitations observed by the experts is a lack of reporting on the similarities between conditions. Ideally, these are the same except for one aspect: the digital game component.

> Apart from the activities implemented, all conditions should be as similar as possible. If you want to examine whether or not learning through a videogame is better than another method, only the game element can be different between the conditions. I definitely think that authors do not report enough on this subject: they do not provide enough information about the conditions, the implementation of the interventions, and how similarity between conditions is attained. In general, I think that the interventions are poorly described. I mean… what do they mean by traditional classroom teaching? I don't know what that is; is it the same as classical classroom teaching? I don't know. For me, it's not the same thing. And isn't this something that is culture dependent? The meaning of traditional classroom teaching will be different in America, Russia, China, etc. So that doesn't say anything.
> (Educational scientist, Professor in instructional psychology and technology, Belgium)

Table 4 gives an overview of aspects on which researchers should try to attain similarity between conditions according to the experts.

Table 4. Aspects of intervention where similarity should be attained in different conditions.

Aspect of intervention	Description
Time exposed	Time exposed to intervention should be exactly the same in both conditions
Content	The exact same learning content should be present in both conditions.
Instructor	The instructor in the different conditions, should be the same person across all conditions.
Support received	Technical support or guidance received by the instructor/intermediaries.
Difficulty level	Content treated in all conditions should be of the same difficulty level
Interaction with other people	Amount if interaction other participants or instructor during the intervention.
Day of the week	Day on which the intervention took place, should be the same in each condition.
Environment	The intervention for all conditions should be conducted in the same environment (e.g. the same classroom)
Types of exercises	Types of exercises used in the game condition should be the same in the other conditions.
Awareness of testing moment	If the game group is briefed about testing moment after the intervention, the other conditions should also be briefed on this.
Reward for participation	If a reward for participation is provided to the game group, this should also be provided to the other conditions.

3.3.3. Instructor

Whilst no preferences regarding the type of instructors present during the intervention were expressed, advantages and disadvantages of the different types were discussed. One disadvantage of adding a non-researcher, such as a teacher, as an instructor is that there is teacher influence: this might impact the results. An advantage of using a researcher as an instructor is that they are trained to give the same instructions to the subjects in each condition in a controlled manner. The presence of a researcher as an instructor can, on the other hand, also lead to bias: people react to changes in environments, for example by aiming to make a good impression. In order to strike a balance between the internal and external validity of a study, experts suggest developing procedures for teacher instruction. In order to control the correct implementation of the procedure, observation by the researcher would be ideal.

In relation to the actual support provided by the instructor, some best practices were mentioned. Whilst some of the experts stated that the most ideal situation would be the absence of an instructor, in order to isolate the game component, even these experts were aware that this is not always possible in this type of research.

> With a good game, the teacher isn't involved at all. In most traditional classroom teaching situations, the teachers are there, at the very least monitoring; offering the odd bit of advice and support. So, it has to do with letting the reader know about the context as a whole.
>
> (Educational scientist, Professor in educational studies and research methods, Scotland)

It is of great importance that the same support is provided in each condition. In some cases, experts prefer that no support is provided at all in order to avoid a confounding effect. Providing procedural help (e.g., support when respondents bump into issues concerning the technology or actual game play or technology oriented support (All, Nuñez Castellar & Van Looy, 2014)) is generally not perceived as problematic: this support would occur even in a formal and even in an informal natural setting (e.g., at home, parents providing this type of help to their children). Consequently, by not providing any support, ecological validity might also be jeopardized. An expert, however, noted that the amount of support that will be provided will be influenced by certain context variables, such as the tech savviness of the parents or teacher or attitude towards games. Hence, it is important to take this into account. Therefore, when it is provided, a clear description is required of what this procedural help actually consisted of.

Guidance (e.g., teacher/supervisor helps to contextualize game play and in game elements in the broader learning context; help related to learning content (All, Nuñez Castellar & Van Looy, 2014)), however, could lead to problems with internal validity. This raises the question of whether or not an effect would have been found if no guidance was provided. Providing guidance also leads to problems with comparability, considering the game condition might need extra help. The absence of guidance, however, could also lead to problems with ecological validity in certain contexts, such as a classroom environment, since asking questions and offering guidance is generally present in this type of environment.

> There should be some rules about that [role of the instructor during the intervention]. Because, if you, as an instructor, think that this intervention requires explanation beyond what you would usually do with a conventional method, this might trigger better learning. Had you given the same explanation to the conventional method, the effect would have been the same, right? So, there have to be rules on how much you reveal and how much you help people because it's a new environment and this level of help and support and assistance has to be comparable across the two treatment conditions. I agree that the educational game treatment might require some more help. But it has to be specified very clearly why and how you assist, why you intervene, and if you do, it would be important to do it for everybody.
> (Experimental psychologist, Researcher on quantitative methodology and experimental methods, UK)

3.3.4. Implementation

One expert believed that DGBL should be implemented as stand-alone intervention. Four experts stated that extra elements could be added to the intervention, as long as these elements are the same in the different conditions. Six experts agreed that the game component should be kept as isolated as possible, considering that other elements added to the intervention could influence learning and thus confound results. Nevertheless, some elements that might be indispensable for practical reasons could be allowed. These elements should, however, be offered in all conditions: for example, even the non-game group should get a training session with the game if this is provided in the game group. The introduction and/or training sessions should not contain substantive information on the learning content covered in the game and should only cover such elements as getting acquainted with the storyline and controls, for example. Otherwise: ‘participants with lower computer skills or less game experience will use up more cognitive capacity in trying to understand the environment,

instead of the game or the idea behind the game' (Experimental psychologist, researcher on quantitative methodology and experimental methods, UK).

Several experts also stated that the necessity to provide procedural help, as discussed in the previous section, might be avoided by providing a training session before the intervention, similar to a trial exercise in psychological experiments. According to more than one third of the experts a debriefing session would not be problematic, if implemented after the post-test, considering a debriefing can entail a learning effect.

If one purely wants to assess the effectiveness of the game itself, providing substantive information regarding the content of the game (e.g., required reading, extra material freely available, game task formulation, etc.) might confound the effect. Therefore, such elements are best excluded from the study.

> Elements such as supplementary material, integration in online platforms, and training sessions result in a confounding effect, caused by extra materials, extra attention, or extra help. This is not the way effectiveness research should be conducted. You should not add extra material if you purely want to assess the effect of the intervention. That is something you should not do. If you want to look at an educational program, of which the game is a part, you can add extra material. But then you should report on the effectiveness of the program and not of the game.
>
> (Educational Scientist, Professor in research methods and statistics, The Netherlands)

Two experts also suggested that, before considering the addition of elements to the digital game, a pilot study on the impact of these different elements of implementation on the learning outcomes could be conducted.

> This has to be pilot tested. If any doubt exists about a training session and the impact it might have on the intervention, that is something you would want to know at the beginning of the study… You can pilot test this, and with this I mean with small groups of 10 to 15 people. Then you execute the intervention without, for example, an introduction and ask the people afterwards whether or not they thought it helped, if they find it necessary, if they had not received the introduction. This way, you have at least some input on that. Whether you eventually use it in your intervention, is up to the researcher, but it is the total package that counts. The effect will be of the intervention as a whole: afterwards you can only speculate about what it would have been without the introduction; but that is another study.
>
> (Experimental Psychologist, Professor in data-analysis and statistics, Belgium)

3.4. *Measures*

3.4.1. Instrument validity

It was generally accepted that when a standardized test on the subject covered in the intervention is available, its use is preferred over one developed for the occasion. The self-developed content could be too closely aligned with the intervention and thus bias the outcomes of the study. It was also recognized that finding a suitable standardized test is not always possible.

When developing a new test it is considered important that the process of development is clearly described: who was involved in the development? Was an expert on the content domain involved? Which factors were taken into account? Et cetera. Moreover, the following elements for reporting on tests used were brought forward by the experts: type of knowledge measured (e.g., knowledge, insight, problem solving, creative ability, etc.), type of questions (e.g., open or closed), difficulty level of the questions, and psychometric properties of the scales used.

Moreover, according to one expert, results should contain average scores without any intervention, scores within certain age categories and a normal distribution of the scores in the target population. Furthermore, several experts favor a 'full disclosure' mentality, adding instruments used if possible. If space is limited, authors should indicate where the test can be found. Finally, several experts suggested that a pilot test would increase validity. This could be done by, for instance, implementing the tests in a similar group of participants, conducting cognition interviews while filling out the tests, examining whether the questions are clear and are interpreted the way they are intended.

With regard to the usage of student achievement (e.g., exam scores) as an indicator of learning outcomes, a majority of the experts (9/13) found it a relevant measure, in terms of ecological validity. However, as student achievement scores can be influenced by other factors, it should be combined with more specific content tests.

> The ultimate goal of serious gaming in a school context would be finding an effect on the school scores (i.e., grades on a test or exam) right? So if you find an effect on your post-measures, but cannot find an effect on school scores, you have a problem. So, for a complete assessment, I think you need both.
> (Experimental psychologist, Professor in research methods and statistics, Belgium)

One expert also indicated that a prerequisite for using student achievement as a measure would be that participants come from the same school. When several schools and educational levels are included in the study, student achievement measures will become difficult to compare. Finally, one expert stated that student achievement only seems relevant for longer-term interventions that have been introduced over a whole semester, for instance.

3.4.2. Similarity pre- and post-test

When using the same test pre- and post-intervention researchers should be cautious about a practice effect. According to three experts, this is less of an issue if the same test is implemented in the other condition(s) as well, considering the primary interest would be the difference between the conditions. Practice effects also depend on the time between the pre- and post-test. A standardized test would be advantageous here, considering these tests typically have guidelines on a minimum amount of time needed between the pre- and post-test.

Using a similar test (e.g., same type and difficulty level of questions) pre- and post-intervention is considered to be a better option by the majority of the experts (8/13), but similarity of both versions needs to be established. Content experts should be involved in the development, but could also serve as 'external assessors'. A better option would be to conduct a pilot study with a group of participants who do not receive the intervention, testing every question separately to see if questions of both tests are matched.

> If you run parallel tests, you should test these beforehand. You would have to match the questions that have the same difficulty level; so that the scores are the same when you implement these tests in two groups. You should also examine the average scores and match questions that in the first version yielded, for instance, 50% correct answers to questions that resulted in 50% correct answers in the second version. And then, you have to make sure that all themes are present in the same proportions in each version. This is not always possible because it implies you would have to conduct a large pilot with the tests you know you will use in your actual experiment later on. Tests should already be created long before your actual study, which is sometimes impossible due to timings and so on. So, if you do not have the luxury of time and you have to create tests that are similar, use some of the same questions in both versions and use some parallel questions that have been evaluated by external assessors on similarity, regarding the type of questions and difficulty level.

(Educational Scientist, Researcher on teaching methods, instruction and assessment, Belgium)

3.5. *Data analysis*

The majority of the experts (10/13) would opt for a standard repeated measures design. This would control for pre-existing differences and take pre-test scores, post-test scores, and comparison of progress across groups into account.

Two experts stated that a mixed effects model should be used, taking both fixed and random effects into account. Fixed effects refer to the effects that are the object of study, which is the instructional condition in this case. Random effects refer to any elements that are observed, and which might lead to extra variance on top of the experimental variance (i.e., the variance between conditions as a result of the difference in treatments, such as different instructors). These are thus potential confounding variables and they should be added as random effects in the model.

> When you conduct analyses, statistics, you have your fixed effects - which you have manipulated - and random effects. In a normal analysis of variance, those random effects could be… you could name thousands, such as sequence, etc. In mixed effect models, those random effects are included in the model.
> (Experimental Psychologist, Professor in psychology, Belgium)

Further, experts emphasized the importance of individual differences in educational research. When aiming to control for certain characteristics, an analysis of covariance is preferably used. Characteristics that are to be examined are added as covariates. Interesting control characteristics suggested by the experts are ability (low, medium and high achievers), computer skills, and previous game experience.

If there is one area where individual differences are very important, it is the learning context. Therefore, the question is not if learning with a game is better than learning without a game. I would preferably put the emphasis on who actually benefits from this game-based approach and who does not. These kinds of individual differences should be taken into account and these variables can be added as a covariate in your analysis. You will notice that there are a lot of differences. This is something you can even do afterwards, without the control group. If you only have the game with its learning effect, you will have people who improve, who worsen, and those where nothing much happens and this could be due to a

variety of elements.

(Experimental Psychologist, Professor in psychology, Belgium)

When differences in pre-test scores are found between conditions, pre-test scores could also be introduced as covariates in order to take these pre-existing differences into account in the analysis.

With regard to reporting, several experts stated that description should not only include significance levels: effect sizes should be reported as well for several reasons. Firstly, reporting on effect size is important to make meta-analyses possible. Secondly, effect size provides more information than significance levels; it also shows how large the effect is. Lastly, by using effect sizes a researcher can estimate how much variance is actually explained by the condition and how much variance is explained by other variables, such as attitude or student achievement.

3.6. *Overview*

Below, on overview of best practices is proposed, based on the expert interviews, which can serve as a guide for the design of future studies (see table 5).

Table 5: Summary of best practices for effectiveness assessment of DGBL (Part 1 of 2).

Best practices	Advantages	Disadvantages
Research design: general		
1. Control group **1.1. *Control group 1: 'Business as usual'***	Control if positive results are not result of mere lapse of time Compare motivational aspects	Larger sample required
1.2. Control group 2: game without educational content (optional)	Compare motivational aspects	Larger sample required
2. Pre-test	• Control for pre-existing differences between experimental and control group(s) • Determine progress/learning gains as a result of the intervention • Make it possible to control for characteristics of drop-outs (i.e., random or non-random attrition)	Practice effects
3. Similarity between experimental and control group should		
a) Randomization of subjects or	Balanced groups in terms of observed and unobserved variables	Larger sample required
b) Randomization of classrooms/schools or	Often more practical in educational research	Classroom/teacher/school influences
c) Matching (blocked randomized design)	Control for similarity between conditions	Unmatched latent variables can influence results
4. Follow-up study (Min. 2 weeks after intervention has finished)	• Control for novelty effect • Control for positive results as a result of intensive training • Control for longer term effects	Attrition
Intervention		
5. Training session	• Might reduce the need for procedural help during the intervention • Less cognitive load is used up for learning to work with the game-environment	Might bias result
6. DGBL as stand-alone intervention	Potential confounds are reduced	Ecological validity might be reduced
6.1. No adding of elements that contain substantive information		
6.2. Instructor role reduced to procedural help		

Table 5: Summary of best practices for effectiveness assessment of DGBL (Part 1 of 2).

7. Instructor type **a) Researcher** **Or**	More experimental control	Ecological validity is jeopardized • Less experimental control
b) Familiar person (current teacher)	Increases ecological validity	• Teacher influences
8. Similarity between interventions should be assured by:	Potential confounds are reduced	Ecological validity might be reduced
8.1. Time of exposure		
8.2. Content		
8.3. Support received		
8.4. Environment		
8.5. Awareness of testing moment		
8.6. Reward for participation		
Participants		
9. Min. 20 participants per condition	Determine differences in dependent variables between groups	More sophisticated analyses are not possible (e.g., covariance adjustment)
Measures		
10. Instrument validity **10.1. Standardized tests** **OR**	• Have been validated • Provide suggestions with regard to timespan required between pre- and post-test for administering the same test	• Might not exactly cover what has been discussed in the game/traditional class
10.2. Ad hoc test developed by researcher	• More closely aligned to content treated in game/traditional class	• Pilot study required • Content might be too closely aligned to content treated in game/class
11. Student achievement (e.g., exam scores)	• Ecological validity measure	• Can be influenced by other factors than the game/control intervention • Pupils should come from the same school (or even class) • Only relevant for longer term interventions (e.g., a whole semester)
Data analysis		
12. Repeated measures	• Analyze interaction between condition and time	• Differences with regard to the dependent variable(s) can exist between groups before the intervention
13. Add random effects if observed	More precise estimate of the treatment effect	Larger sample size required
14. Add participant characteristics as covariates (e.g., game experience, computer skills, ability)	Take into account individual differences in order to determine for whom DGBL interventions are more beneficial	Larger sample size required

4. Discussion

The present study aimed to define best practices for assessing the effectiveness of DGBL, based on interviews with experts in pedagogy and psychology.

When we compare our results to current practices in DGBL effectiveness assessment, some areas can definitely be improved. Firstly, follow-up studies are rarely implemented (Backlund and Hendrix 2013). They are considered important in the case of DGBL due to the 'novelty' of the gaming medium (Clark 2007). Follow up studies help to establish if short-term beneficial effects might be an overestimation of the instructional effect. Secondly, the role of the instructor should be reduced to procedural help in order to define the effectiveness of the game as such. This conflicts with current DGBL literature, where the instructor and the creation of a meaningful learning context are considered key elements in achieving DGBL effectiveness (Giessen 2015, Cruz et al. 2015; De Freitas, 2006). Thirdly, a large part of the present studies have (unintentionally) added confounding elements to the intervention (All, Nuñez Castellar & Van Looy, 2014), making it impossible to know whether or not an effect would have taken place without these elements. Future studies should aim to isolate the game as much as possible and provide more information on characteristics of the implementation of interventions. Lastly, more information should be provided on efforts to achieve similarity between experimental and control conditions, in order for readers to judge the quality of the papers.

The present manuscript confirms several general best practices in experimental research on educational interventions that should also be implemented in DGBL effectiveness research. However, it also brings forward several best practices specific for experimental research in a DGBL context. For example, individual characteristics/differences are more important compared to 'general effectiveness evaluations' when conducting DGBL research. Previous game-experience and computer skills might also influence the effectiveness of the DGBL intervention. Hence, these variables are important to take into account and control for when assigning participants to conditions.

Beyond the characteristics of the learners themselves, the characteristics of the environment in which DGBL is implemented are relevant. For instance, if implemented among children at home, tech savviness or parental gaming experience can play an important role, as it might lead to increased procedural help (help related to issues regarding technology or game play). If implemented in a school context, tech savviness of the teacher and his/her attitude towards digital games as an instructional tool could also influence results, and this should be

taken into account. Moreover, DGBL effectiveness studies in a lab context should be avoided, due to the important role of intrinsic motivation/enjoyment plays in the learning process. This is why it is important - besides the general good practice reasons - that a control group receiving another educational activity is added to the study design (i.e., in order to compare motivational outcomes). This is in line with recently published research, where higher scores on motivational outcomes in the DGBL intervention can be a decisive factor for implementing/adopting DGBL, even if similar cognitive learning outcomes are achieved (All, Nuñez Castellar & Van Looy, 2015). DGBL effectiveness studies can - depending on the aim of the game, such as practicing math at home - be conducted in a context where the researcher does not have much control. Hence, other control mechanisms specific to the game environment need to be implemented, such as gathering log data on gameplay. Even more than in general effectiveness evaluation research is the addition of a follow-up study, in order to control for a short-term novelty effect.

An area of tension that clearly remains is the issue of internal versus external validity of studies. It is clear from the results that DGBL effectiveness research cannot be conducted using the Fisher tradition (Fisher 1934, 1935), because the researcher cannot have complete control over the experiment. This is caused by the complexity of the environments where games are typically implemented, such as implementation of the interventions in natural collectives (e.g., existing class groups). Additionally, different unobserved variables can influence the outcome results and implementation in different contexts (e.g., games targeted to play at home, field experiments). Results of this study show that a proper balance between internal and external validity is best achieved by improving both aspects. Internal validity can be increased by reducing the influence of confounding variable during implementation of the intervention(s) and by adjusting analysis for potential sources of variability other than the experimental variance. External validity can be maximized by ensuring similarity between elements present in the real world implementation environment and the implementation for the effectiveness assessment, such as implementation in a context in which the game is intended to be used, implementation in natural collectives such as existing class groups (i.e., randomization on a classroom level), the presence of a familiar teacher in a classroom, the provision of procedural help, et cetera.

Several suggested best practices, however, need to be further evaluated with regard to DGBL. For instance, keeping time exposed to intervention equal in the experimental and control group might seem like a general good practice. However, in DGBL, this idea is not as straightforward because playtime does not equal time that is spent on the learning content. Moreover, an important desired outcome of implementing DGBL is cost-efficiency, of which

improved time management (i.e., a reduction in time spent to learn a certain content matter) is an important indicator (All, Nuñez Castellar & Van Looy, 2015). When keeping time spent on the intervention equal, this prevents researchers from testing if DGBL scores better on time management. Another element, on which similarity between conditions might not be straightforward, is the reduction of support to procedural help and thus leaving out any help providing substantive information. This might jeopardize ecological validity for the control group when this is a traditional class, for instance, where asking questions related to the learning content is a common practice. Also, by providing the control group with a game training session in order to safeguard similarity between conditions, the psychological impact (e.g., in relation to motivation) of preparing participants for a game they do not get to play is ignored, which in turn, can be a confounding factor.

In conclusion, it is clear that a more standardized approach is not only possible but required in order to be able to improve rigorousness of DGBL effectiveness research and define guidelines. For instance, in order to be able to conduct a power analysis to determine which sample size one needs for his/her research, assumptions on the magnitude of the effect of the DGBL intervention need to be made. However, this assumption on the magnitude of the effect is a difficult exercise for researchers to make, as in one study the control group did not receive any educational intervention and in another study the control group got instructed using a 'classical' approach. Elements like this limit the field to grow from evaluation into research, which requires more rigorous standards of validity and reliability (Hutchinson 1999). With this study we hope to have taken an important step in this direction.

5. Future Research & Limitations

Future research should focus on the development of a research protocol that supports DGBL effectiveness research. The present study took a first step in this direction by defining best practices. An interesting venue for further research would be to compare results of different study designs in published and unpublished (in order to account for publication bias) DGBL effectiveness research against our proposed best practices.

A limitation of the present study is that we cannot make claims about the representativeness of these results due to the limited number of participants and potential geographical bias. We have aimed to achieve acceptable coverage by including both national and international experts and by gathering data until saturation was achieved. An interesting

approach to improve representativeness, would be conducting a survey study among a larger sample of expert in order to quantitatively validate our results.

The present study focuses on a research design for a summative evaluation and aims to determine if a DGBL intervention is successful or not. It does not provide any indication on why an intervention worked or did not work. The addition of formative evaluation research for determining what makes a DGBL intervention effective is recommended. The present study has also not taken into account the growing area of stealth assessment (i.e., unobtrusively gathering in game information as indicators for learning and motivation). Future research should definitely focus on combining results of effectiveness' studies using an experimental design and the results from stealth assessment.

Lastly, the best practices brought forward are based on the experience and expertise of academic researchers, suggesting an ideal design. In practice, however, some issues may occur which prohibit the implementation of the suggested recommendations. Hence, the authors aim to validate these recommendations by implementing them in DGBL effectiveness studies in different contexts (e.g., school, workplace, health context). This way, the recommendations can be optimized in order to develop a standardized procedure for assessing the effectiveness of DGBL that can be flexibly used across different sectors

PART 2.

FEASABILITY STUDIES

CHAPTER 5.
Learning English in primary school: comparing short and long term effects of a game-based and traditional classroom approach.

Abstract

Digital game-based learning (DGBL) is increasingly implemented in a variety of sectors including education, industry and health. While a lot of research has been conducted regarding its effectiveness, certain authors have pointed out elements that jeopardize reliability and validity of some findings. Beneficial effects of DGBL have for example been attributed to a potential novelty effect. Moreover, lack of strict experimental controls have triggered the criticism that effects of secondary learning interventions such as a debriefing session may confound results.

In order to test these hypotheses, a randomized controlled trial was conducted with seventy-one primary school children of the fourth and fifth grade. One group received traditional classroom instruction, a second group played an educational game and a third group played the game, followed by a debriefing. Importantly, the participants were exposed to exactly the same English vocabulary content for an equal amount of time, and the outcomes of the three different types of training were assessed longitudinally by means of a second post-test after three weeks.

Results show a mixed picture with the game-based intervention providing more enjoyment but similar learning effects in the short term and a lower retention score in the longer term compared to the traditional learning condition. Moreover, in the game with debriefing condition, participants reported similar enjoyment as in the game condition but displayed lower learning than in the traditional learning condition both in the short term. This effect may be due to the strict control of time in the setup whereby the implementation of the debriefing session led to lower game time and thus less exposure to the learning materials. In conclusion, our results indicate that DGBL is effective in terms of affective outcomes such as enjoyment yet may yield lower long-term learning effects than traditional teaching.

Keywords:

elementary education; evaluation methodologies; evaluation of CAL systems; interactive learning environments

Reference:

All, A., Nunez Castellar, E. P., Meesschaert, D., & Van Looy, J.(in review). Learning English in primary school: comparing short and long term effects of a game-based and traditional classroom approach.

Preliminary results presented at the *the 65th ICA annual conference 'Communicating with Power', San Juan, Puerto Rico*: All, A., Nunez Castellar, E. P., Meesschaert, D., & Van Looy, J. (2015). Validating a standardized procedure for effectiveness assessment: learning English vocabulary through gameplay.

1. Introduction

The interest in using digital games as instructional tools has grown exponentially over the past decade. digital game-based learning (DGBL) has been implemented in a wide variety of sectors including defense, education, corporate training, health and wellbeing, and communication (Backlund & Hendrix, 2013). DGBL refers to an instructional medium that uses the entertaining power of games to serve an educational purpose (Prensky 2001). Digital games have the power to intrinsically motivate players to engage in the activity (i.e., performing the activity in itself and for itself (Ryan & Deci, 2000a)), which has been considered an important aspect of games to boost learning (Garris, Ahlers, & Driskell, 2002). More specifically, intrinsic motivation for performing an activity is associated with higher levels of enjoyment, interest, performance, higher quality of learning and a heightened self-esteem (Ryan & Deci, 2000b). Successful DGBL is thus the result of a balance between the learning and the gaming element (Nussbaum & Beserra, 2014). This implies that the goal of DGBL is twofold: it has to be entertaining and it has to be educational (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013). Consequently, in DGBL literature and published effectiveness studies both learning and player engagement/motivation are considered relevant to assess (Bellotti et al. 2013).

The increased interest in and implementation of DGBL has resulted in a need to know whether or not these instructional tools are effective (Mayer, et al., 2014) and whether the substantial financial effort required to develop and implement DGBL is worthwhile (Clark, 2007). Concomitantly, there is a growing need to conduct empirical studies aimed to assess the effectiveness of DGBL (Bellotti, et al., 2013; Mayer, et al., 2014). In recent years, a significant amount of research assessing the effectiveness of DGBL has been published (Hainey, et al., 2014; Hwang & Wu, 2012). Results regarding its effectiveness are, however, mixed. While some meta-analyses have found that DGBL is more effective than non-game instructional methods regarding learning gains, others have found non-significant differences (Backlund & Hendrix, 2013; Clark, Tanner-Smith, & Killingsworth, 2015). The same inconsistency is found regarding motivational outcomes (Clark, et al., 2016; Wouters, Van Nimwegen, Van Oostendorp, & Van Der Spek, 2013)

Apart from the diverging nature of game genres and topics that comprise DGBL, heterogeneity in study designs is one of the major reasons why effectiveness studies yield mixed results and why results cannot be compared across studies (All, Nuñez Castellar & Van Looy, 2014). A big issue is the addition of extra elements to DGBL interventions such as required

reading, a debriefing session and extra exercises. This makes it difficult to know whether the effect would have occurred without these elements as previous research has shown that providing additional non-game instruction enhances learning outcomes compared to DGBL where no additional non-game instruction was provided (Clark, et al., 2016; Sitzmann, 2011; Wouters, et al., 2013). Moreover, in the last decade critiques have been formulated regarding to the rigorousness of effectiveness studies on DGBL. Some main issues are comparisons with control groups that did not receive an educational intervention (Hays, 2005) and time-on-task differences between experimental and control groups (Clark, 2007). Moreover, a systematic literature review conducted by the authors brought forward some suboptimal study design elements related to confounds. Firstly, the addition of elements to the game, such as required reading, extra exercises, or debriefing sessions, makes it impossible to isolate the effect of the game. Secondly, the role the instructor has during the intervention differs across studies and can possibly influence results. Instructors are either present to 1) only supervise, 2) offer technology oriented support when respondents encounter issues concerning the technology or actual game play (i.e., procedural help) or 3) offer content-related help, by providing contextualization of game play and in game elements in the broader learning context during actual game play (i.e., guidance) (All, Nuñez Castellar & Van Looy, 2014). Help involving substantive information should, however, be avoided, as this can potentially influence learning outcomes (All, Nuñez Castellar & Van Looy, 2016). Thirdly, implementation of the same test pre- and post-intervention on the same day, could lead to practice effects and pre-test sensitization. This would, again, result in an overestimation of the instructional effect (Crawford, Stewart &Moore, 1989; Van Engelenburg, 1999). Hence, preferably parallel tests are used pre- and post-intervention to reduce these effects (All, Nuñez Castellar & Van Looy, 2016) . In 1992, Randel mentioned similar issues with regards to the reliability and validity of certain effectiveness studies on instructional games. Twenty years later, the same issues are still detected in DGBL effectiveness research.

1.1. *Debriefing*

Debriefing in educational contexts, refers to methods used in experiential learning models to stimulate post-experience reflection on and analysis of the content treated in the intervention (Fanning & Gaba, 2007; Nicholson, 2012). The stimulation of reflection can also be incorporated in the game by means of scaffolding techniques, referring to assistance/support during the game experience to gain better levels of understanding of the learning content in the

game. These can for instance be question/reflection prompts popping up when the player executes a certain action (Adams & Clark, 2014; Chen & Law, 2016; Vrugte et al., 2015). In the present study we focus on post-experience reflection and analysis.

At the theoretical level, debriefing has been considered as a potentially important element in DGBL in order to connect the game experience to learning (Crookall, 2014; Pivec, 2007; Van Der Meij, Leemkuil, & Li, 2013) and to enhance lasting learning effects and transfer (Kriz, 2008). Moreover, it has been speculated that the absence of a debriefing session could result in a negative game experience, as it can leave learners upset, confused or angry because they might not know what it was all for (Crookall, 2014). A debriefing can help reduce negative feelings about aspects of the activity (Nicholson, 2012).

Different types of debriefing methods can be distinguished. Firstly, there are meaningful differences between an expert-led debriefing, referring to a facilitator present leading the debriefing or a self-debriefing, referring to debriefing among teams of learners or individual debriefing. Secondly, one has to choose between an oral and a written debriefing and lastly, between an individual or a collaborative debriefing (Van Der Meij, et al., 2013). A debriefing typically consists of several phases of which the number differs depending on the model used. Many of these models are rooted in Kolb's experiential learning cycle (Fanning & Gaba, 2007), which proposes four phases for a debriefing. The first phase 'concrete experience' refers to the learning activity itself. The second stage 'reflective observation' refers to the conscious reflection on the activity, by reviewing what one has done and experienced. The third phase 'abstract conceptualization' refers to making sense of what happened during the activity and thus linking it to intended learning outcomes, interpreting events and relationships between them. The fourth and final phase is that of 'active experimentation', reflecting on how the content learned can be applied in real life situations (Kolb, 1984). Another important element regarding debriefing is time spent. As a minimal requirement, time spent on the debriefing should at least be equal to the time spent on the actual intervention (Boud, Keogh, & Walker, 2013). Van Der Meij and colleagues (2013) have provided a conceptual framework for debriefing in a DGBL context, adding topics discussed and leading questions (table 1).

Table 1: Conceptual model for debriefing with phases, topics and leading questions

Phases	Topics	Leading question(s)
Concrete experience	Events	What happened?
	Emotions	How did you feel?
Reflective observation	Empathy	How do you value this experience?
Abstract conceptualization	Explanations	What did you learn? What would have happened if . . . ?
Active experimentation	Every day	How are the game events and reality connected?
	Employment	How do you go on from here?
	Evaluation	What would you do differently?

While the addition of a debriefing session is considered by many as a potentially important element in a DGBL context, empirical research on the effectiveness of debriefing in a DGBL context is scarce. A recent study conducted by Bakker and colleagues (2015) has compared playing math mini-games at home (home), to playing at home and receiving a debriefing at school (home-school condition) and playing at school followed by a debriefing (school condition). In this study, the home-school condition was most effective regarding learning outcomes; no differences between the home and school condition were found. Playtime was not controlled in this study. The games were played most frequently in the school condition and least frequently in home condition, with all between-condition differences being significant. Hence, the authors concluded that a debriefing session in school did add value for both learning outcomes and engagement, as the home-school condition spent significantly more time playing the games at home. Another study investigated the added value of clarifying concepts present in a DGBL context (i.e., business concepts such as cost, price, profit) before (pre-briefing) or after (debriefing) game play, compared to a condition in which participants only played the game (Barzilai & Blau, 2014). However, the participants of the pre- and debriefing condition spent additional time with the learning content, which is a potential confound when compared to the only-play condition. Results showed that participants who received clarification of the concepts before gameplay outperformed the other two groups. Hence, the authors concluded that debriefing might not be sufficient for helping learners to form connections between the game and what they learn at school. This is in contrast with the statements made in literature regarding the potential added value of debriefing and learning through games. Also, no differences were found between groups regarding game experience (i.e., enjoyment). Similar mixed results are found for the in-game scaffolding techniques stimulating reflection (Adams & Clark, 2014; Chen & Law,

2016; Vrugte et al., 2015). However in one case where it did not add value, this was related to differences in progress made in the game as a result of these in-game question prompts (Adams & Clark, 2014).

Considering the mixed results, the different gameplay contexts and differences in time-on task in these studies, we aim to test the added value of a debriefing in a DGBL intervention implemented in class, whilst strictly controlling for time and content.

1.2. *Long term assessment of DGBL effectiveness*

A common finding when reviewing the DGBL literature, is that follow-up tests to assess lasting effects of DGBL are rarely implemented (Hauge, et al., 2014). Implementing a follow-up test, however, can be beneficial for several reasons. Firstly, positive results in favor of DGBL on a first post-test could be the result of a novelty effect, creating a short-term effect which might not be sustained over time (Galarneau, 2005) and thus overestimating its instructional value (Clark, 2007). For instance, a meta-analysis of Clark (1985) aimed at exploring internal and external validity of effectiveness studies in technology delivered instruction, found that the strongest effects are found in short-term studies, where post-tests are introduced shortly after the instruction. Secondly, positive results could be due to intensive training, considering post-tests in DGBL effectiveness studies are often introduced directly after the intervention (All, Nuñez Castellar & Van Looy, 2016). Thirdly, research has shown that, despite the fact that no immediate effect is found in the first post-test after the intervention, effects can be found in the follow-up tests (Brom, Preuss, & Klement, 2011; Randel, et al., 1992). Finally, by combining a follow-up test and a debriefing condition whilst controlling for time and content, the present study will allow us to examine whether a debriefing contributes to lasting learning effects of DGBL (Kriz, 2008).

2. Aim of the study.

The current study is part of a larger project that aims at formulating guidelines for conducting effectiveness studies of DGBL by means of a standardized procedure. Two phases have been conceived in this project: a first phase consisted of the development of the procedure and the second phase consists of validation by conducting experiments in different contexts where DGBL is frequently being used as an instructional method such as health, educational and corporate contexts (Michaud, Alvarez, Alvarez, & Djaouti, 2012). Based on a systematic

literature review we have mapped current methods used for conducting effectiveness of DGBL (All, Nuñez Castellar & Van Looy, 2014) and based on these results and expert interviews we have defined best practices for conducting effectiveness studies on DGBL (All, Nuñez Castellar & Van Looy, 2016). By means of a user requirements analysis, we have also operationalized the concept of effectiveness in a DGBL context (All, Nuñez Castellar & Van Looy, 2015). Based on these previous steps, a first version of the standardized procedure was developed.

In this study we present the first of the validation studies conducted in an educational context. More specifically, we aim to test the effectiveness of an English language learning game platform for primary school children. On the one hand we aimed to test the feasibility of the procedure and to map potential barriers to its implementation by conducting an effectiveness study using the procedure. While the procedure consists of guidelines for every aspect of the study design and due to space limitations, we cannot discuss them all. The checklist of the procedure can be found in appendix E. Examples of elements we want to test in these validation studies are the extent to which comparability between experimental and control groups can be assured, to what extent randomization of assignment of participants to conditions is possible, to what extent potential confounds can be reduced –in this case- in a classroom environment and when a follow-up study should be implemented.

However, the specific aim of this study and the focus of this paper was to investigate the impact of two factors that have been under discussion in relation to DGBL effectiveness. The first element that we investigated was whether the claimed beneficial effects of DGBL are maintained on the longer term. In other words, whether a second post-test could help researchers gain insight into the long-term impact of this instructional method. The second element we aimed to assess was to what extent the inclusion of a debriefing session can influence DGBL learning outcomes. Although from the pedagogical point of view, a debriefing is considered of added value to an educational intervention (Crookall, 2014; Garris, et al., 2002), it often entails additional time spent with the actual learning material which is a potential confound for the assessment of DGBL interventions (All, Nuñez Castellar & Van Looy, 2014). Therefore, in the present study we included an intervention in which we incorporated a debriefing session while controlling for instructional time.

2.1. *Hypotheses*

DGBL is considered effective with regard to learning outcomes if it succeeds in achieving at least similar learning gains compared to standard teaching practices (All, Nuñez Castellar &

Van Looy, 2015), without significantly diminishing motivational and efficiency outcomes. In this case, we compare a DGBL platform for teaching English as a second language to a class which is given by the teacher (see section 3.3.2. for more detailed information on the traditional class). Hence, we hypothesize that in the short term (post-test 1), no significant differences will be observed with regard to vocabulary knowledge between the game group and the traditional classroom group (H1). If similar learning gains are achieved compared to the current method, a decisive factor for adopting and implementing DGBL is higher levels of motivational outcomes (All, Nuñez Castellar & Van Looy, 2015). Hence, our second hypothesis is that the game groups will score significantly higher on enjoyment in comparison to the traditional classroom group (H2). Considering that the DGBL platform is based on experiential learning, we also hypothesize that pupils receiving a debriefing session after the DGBL intervention will outperform the DGBL group that did not receive a debriefing on the English vocabulary test (H1a) and will score significantly higher on enjoyment compared to the group that only played the game (H2a). Regarding longer term impact, we hypothesize that in order for the game to be considered effective, there will be no differences on the follow-up test on English vocabulary between the game and the traditional class condition (H3). Based on the debriefing literature, we also hypothesize that regarding enjoyment, the game + debrief group will outperform the game condition that did not receive a debriefing on the follow-up test on English vocabulary and consequently, the traditional class (H3a).

3. Method

3.1. *Design*

Children of the fourth and fifth grade of primary school were randomly assigned to either a game condition, a game + debriefing condition or a traditional class condition. Children in the game condition were taught English vocabulary using a DGBL platform; the game + debriefing condition were instructed using the same platform, but received a debriefing session afterwards and in the traditional classroom condition children were instructed by the teacher, by means of a traditional class. There were three test sessions in the present study: before the intervention, after the intervention and a follow-up test 3 weeks after the intervention was completed (see table 2). Importantly, we strictly controlled that the time and content were exactly the same in the three conditions.

Table 2: Overview of procedure followed in the 3 conditions.

Condition	Pre-test	Game	Debrief	Class	Enjoyment scale + Post-test 1	Post-test 2	Total time	Instructor	Role instructor
Game	15 min	40 min	/	/	5 + 15 min	15 min	90 min	ICT Coördinator & Researcher	Procedural help
Game + debriefing	15 min	20 min	20 min	/	5 + 15 min	15 min	90 min	Teacher, ICT Coördinator & Researcher	Procedural help & leading debriefing session
Traditional class	15 min	/	/	40	5 + 15 min	15 min	90 min	Teacher	Content instruction

3.2.*Participants*

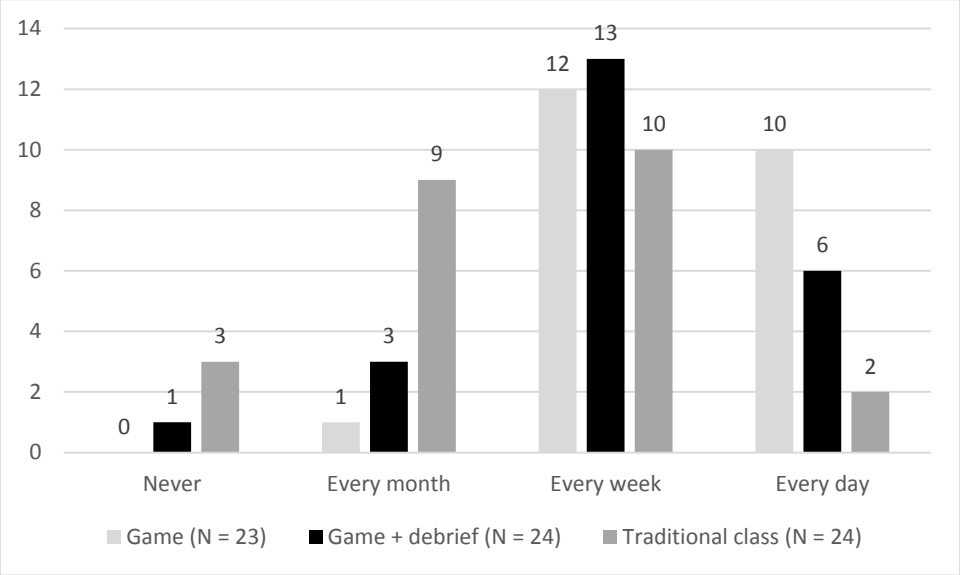
This study was conducted in collaboration with a primary school in Flanders. Primary school children do not learn English in school, but might be in the possession of a limited knowledge of the English language due to media exposure. In Flanders, TV-shows are often in English and subtitled. Hence, our results might not carry over to other countries or even the French speaking part of Belgium, with less media-exposure to the English language. Pupils of the fourth and fifth grade participated in the study. Each grade consisted of 2 class groups. Randomization of subjects over each condition was used in order to avoid influences of age, gender, previous English vocabulary knowledge and game experience. All pupils that had the class group number (i.e., a number is assigned to the pupils at the beginning of the school year based on an alphabetical ranking) 1 to 6 were assigned to the game condition, all pupils that had the class group number 7-12 were assigned to the game + debriefing condition and all pupils that had the class group number 13-18 were assigned to the traditional class condition. In total, 72 subjects participated in the study. However, 1 pupil was absent the day of the intervention, thus the data reported in the current study includes 71 subjects. As can be seen in table 3, no significant differences can be found between the experimental conditions in terms of gender distribution,

previous game experience or previous knowledge of English vocabulary or grade the participants belong to. Previous game experience was assessed by asking the participants how frequently they play games (referring to every type of game genre and every type of platform, such as a PlayStation, xBox, the computer, PSP, mobile phone, tablet, etc.) with response categories 'never' , 'every month', 'every week' or 'every day'. Unfortunately, the assumptions for chi square analysis were not met with the original categorization of answers. Hence, we have recoded this variable into a dummy variable, defining gamers as participants who play games at least once a month. We have also added a chart to provide insight in the responses of the participants on the original question.

Table 3: Control for balanced groups as a result of randomization on subject level.

	Educational game (N =23)	**Educational game + debriefing session (N = 24)**	**Traditional class (N = 24)**	**Chi²**	**p**
Female gender	n = 17	n= 16	n= 14	1.28	.53
Gamers	n = 23	n = 21	n = 22	3.53	.17
Grade (fourth/fifth grade)	n = 12/n = 11	n = 11/ n = 13	n = 12/ n =12	.20	.91
				F	**p**
Pre-test score (/20)	M = 9.04 SD = 4.72	M = 8.27 SD = 4.33	M = 7.48 SD = 4.01	.757	.47

Figure 2: Game frequency per condition*



***** one case in the game + debrief failed to fill out the gaming frequency question

3.3.*Stimulus material*

3.3.1. Digital game-based learning platform

For the present study, –removed for blind peer review process- , a digital game-based learning platform was used, which aims at teaching English. We have chosen for the school edition aimed at 6 to 12 year olds. The DGBL learning platform can be played on tablets or on a computer in a web browser. For the current study, the browser-based version was used and logins were provided by the developer. The game is a virtual world called - blinded for review purposes- where the –name removed for blind review purposes- family lives. More information about the game can be found here:–website removed for blind review process-. The game itself consists of 10 thematic missions. Every mission consists of several activities where they can gather coins for correct answers. Once an activity is successfully completed, the pupil earns a 'feather' and once a pupil has gathered 3 feathers, he/she reaches another level. The activities can either be a) educational games b) interactive stories and c) creative lab activities. Every mission starts with an interactive story to let the pupils get acquainted with the vocabulary to be practiced in the mission. <Name DGBL platform removed for blind review purposes> can thus be considered as an extrinsic game type (Ang & Zaphiris, 2006), i.e. "a structured series of puzzles or tasks embedded in a game or narrative structure with which they have only the most slender connection" (p. 10).

Subjects belonging to the game condition and the game + debriefing condition, could freely play the first six activities of the first mission 'the family' where English vocabulary with regard to family members and colors are discussed. In mission 1, the family is to take the family photo, but little sister Andrea is missing. The family photo cannot be taken before she is found. An overview of the activities can be found in table 4.

Table 4: Activities of mission 1 of Mingoville pupils had to complete for the study.

Activity	Type	Description
Meet my family	Interactive story	Recognizing and indicating Candy's family members
Color my family	Game	Coloring family members in the colors instructed
Play Memory	Game	Linking images to written words, spoken words to written words and spoken words to images in the form of a memory game
When Ryan met Martha	Interactive story	Listening exercise accompanied by text, where the students need to illustrate the story by linking the images to the storyline
Make a family picture	Creative lab	Kids have to make their own picture by dragging the picture elements from the mission on to the screen.
Pacman	Game	Spelling words by eating the correct letters in the context of a pacman game

3.3.2. Traditional classroom materials

The teacher was provided with a lesson schedule to follow when giving the class. The lesson was designed in such a way, that it covered exactly the same content that was treated in the game. For this, a content analysis of the DGBL platform was conducted. A large drawing with family members was also provided to be presented on the blackboard, to keep the instructional method as similar as possible. Also, paper and pencil exercises were provided for the pupils, which served as support material during class and were jointly filled out. This was added to the lesson, because this is the way pupils are used to being instructed in the classroom and deviation from the current way the lessons are designed was minimized. The lesson plan, the blackboard drawing and the paper and pencil exercises can be found in appendix A.

3.4.*Measures*

A user requirements analysis conducted by the authors (All, Nuñez castellar & Van Looy, 2015) suggested that DGBL effectiveness consists of three categories: learning, motivational

and efficiency outcomes for which several indicators can be used (see table 5). At least one learning outcome and one motivational outcome should be assessed as a minimal requirement for effectiveness assessment of DGBL as its goal is twofold: instruction and entertainment. Hence, we chose to assess objective performance by means of a vocabulary test (learning outcome) and enjoyment by means of the Relative Enjoyment Scale (motivational outcome).

Table 5: Operationalization of DGBL effectiveness

Learning outcomes	**Motivational outcomes**	**Efficiency outcomes**
Situational interest DGBL stimulates interest in the content matter discussed in the game.	***Enjoyment*** DGBL succeeds in creating an enjoyable game experience.	***Time management*** DGBL succeeds in reducing the timeframe required to teach a certain content matter. This is a judgment of relative worth compared to other instructional methods.
Performance DGBL succeeds achieving learning goals as defined by the game developer/the client who ordered the development of the game.	***Motivation towards DGBL*** Learning with the game-based method is motivating. This is a judgment of relative worth, compared to other instructional methods.	***Cost-effectiveness*** DGBL succeeds in reducing the cost of the intervention with regard to: a) the number of learners that can be reached and b) the time required to teach the target group certain content. This is a judgment of relative worth, compared to other instructional methods.
Transfer DGBL stimulates application of learned content matter in the game to real world situations.		

Adapted from All, A., Nuñez Castellar, E.P. & Van Looy, J. (2016). Towards a conceptual framework for assessing the effectiveness of digital game-based learning. Computers & Education, 88, 29–37.

3.4.1. English vocabulary test

Considering that English is not part of the school curriculum for the pupils who participated in this study, validated English tests were not available at the moment of testing. Hence, a test was developed by the authors in cooperation with one of the teachers based on an analysis of the game content. Three tests were developed: a pre-test, a post-test and a second post-test which was administered 3 weeks after the end of the intervention. The pre- and post-tests were not identical, but fully interchangeable tests (i.e., same types of questions and same type of difficulty level), in order to reduce the risks of a practice effect, considering that the pre-test and post-test were administered the same day. The pre- and first post-test were piloted (N = 15) to see if undesirable effects could be observed (i.e., ceiling effect, non-normal

distribution of the data or non-parallelism of the pre- and post-test). An analysis of variance indicated that the pre- test (M = 16.37, SD = 1.59) and post-test (M = 15.69, SD = 2.12) could be considered as interchangeable versions (F =.49, p = .50). Also, we failed to find deviations from normality and homogeneity of variances. The follow-up was developed according to the exact same protocol as the pre- and the first post-test. All tests can be found in appendix B.

3.4.2. Enjoyment

To assess enjoyment, a preliminary version of the Relative Enjoyment Scale (RES) was used (Van Looy, 2016). We chose this scale, because it previously has failed to find deviations from normality (Van Looy, Nuñez Castellar & Houttekiet, 2016). The scale consists of 12 items, whereby the subjects have to compare the target activity (educational game/lesson) with other activities, such as swimming or riding a bicycle on a seven point Likert scale. An example page can be found in appendix C. The RES appeared to have good internal consistency for our data ($\alpha = 0.80$).

3.5.*Procedure*

The interventions were all implemented on the same day at the same primary school. Both the game and traditional classroom condition received the intervention during the first hour of lessons (9h-10h15); the game + debriefing condition received the intervention the second hour of lessons (10h15-11h30). A schematic overview of the procedure can be found in table 1 in section 3.1.

In the game conditions, the computers were already turned on and a shortcut to –removed for blind peer review process- was provided on the desktop. Every computer was fitted with a note containing a login and a password. The pupils only received a short introduction on how the game worked. The game was played individually in the computer classroom of the school. Considering that the game has different auditory elements, pupils were asked beforehand to bring headphones. During the intervention, a researcher and the teacher from the school were present. Only procedural help (i.e., related to problems with gameplay or the computer) was provided, meaning substantive information related to the actual content of the game was withheld. After 20 minutes, the pupils in the game + debrief condition were requested to end the game because 20 minutes of debriefing would follow. We have chosen for an expert-led oral debriefing session, which was structured according to the conceptual model proposed by Van Der Meij and colleagues (2013) presented in table 1. The debriefing session was led by the

teacher and in appendix D, notes taken during the debriefing can be found. In the game condition without debriefing, the pupils could continue playing the first mission of -name game removed for blind review process-. The third condition received a traditional class for 40 minutes. During this intervention, only the teacher was present. Due to practical limitations (the researcher was present in the game condition), no researcher could be present. In appendix B the procedure of the class is described in detail (see a) Lesson plan).

3.6.*Data Analyses*

Data were analyzed using SPSS 22.0 statistics software. First, we conducted chi square tests to check for differences between groups regarding gender and game experience. A next step was to test whether there were pre-existing differences between groups with regard to English vocabulary knowledge by conducting an analysis of variance (ANOVA) on the pre-test scores. A repeated measures MANOVA was conducted to determine the effect of testing moments, treatment and whether differences between testing moments differ across conditions. Additionally, new variables were introduced in order to compare progress between groups (i.e., difference in difference method using analysis of variance, see Gerber & Green (2012). The post-hoc test for unplanned comparisons (Scheffé) was used for the English vocabulary test, as we wanted to compare all instructional methods.

The relative enjoyment scale consisted of 2 sheets of paper, printed on both sides. On each page, 3 items of the relative enjoyment scale were displayed. One pupil did not fill in the relative enjoyment scale and was excluded from the analysis of the RES scores. Eight other pupils skipped a complete page and thus 3 of the 12 items (one pupil did not fill in the first page, one did not fill in the second page and 5 did not fill in the third page). Hence, we can define the missing data for the relative enjoyment scale as item nonresponse referring to nonresponse on particular items in the questionnaire (Little & Rubin, 2002). Hence, we can consider our missing data being randomly missing as the missingness is related to a certain page on which they were presented. Since the data is not missing completely at random and more than 5% of certain items on the Relative Enjoyment Scale are missing, we cannot simply conduct our analysis only with the respondents of which we have complete data of the RES (Little & Rubin, 2002; Scheffer, 2002). Hence, we have used single imputation using the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977) using Missing Values Analysis within SPSS 22.0 to deal with the missing data.

Effect size r was calculated to estimate the size of the effects. Effect size r for t-test analyses was calculated by taking the square root of $t^2/(t^2+df)$. For analyses of variance, effect size was calculated by taking the square root of eta squared which was obtained using the following formula: $SS_{between}/SS_{total}$. For the non-parametric Kruskal-Wallis test, effect size r was calculated by taking the square root of the eta squared which was calculated using the following formula: $(H - k + 1)/(n - k)$, where k equals the number of groups in the study.

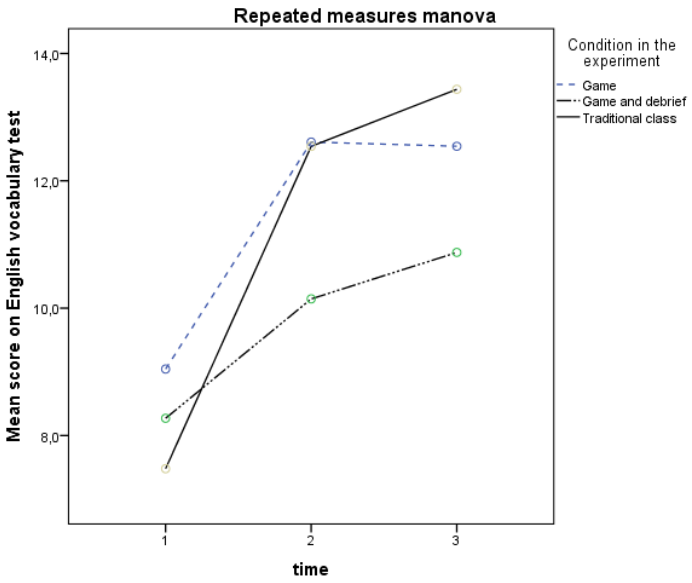
Before analyses of variance were conducted, we checked whether assumptions underlying analysis of variance were met. We checked for homogeneity of variance using Levene's test and a scatterplot with the predicted values and studentized residuals of the dependent variable. We checked for normality, using the Kolmogrov-Smirnov test, by looking at skewness, kurtosis and a Q-Q plot of standardized residuals of the dependent variable. These statistical tests failed to find deviations from normality. eThe assumption on homogeneity of variance was not met for the RES. Consequently, we have checked for outliers using the median absolute deviation, which is considered a robust method for detecting outliers, as it is not influenced by either sample size or the value of the mean (Leys, Ley, Klein, Bernard, & Licata, 2013). One extreme low value of 24 was (i.e., +/- 2.5 median absolute deviation) present in the data for the RES. As ANOVA is quite sensitive to outliers (McClelland, 2000), we have conducted the non-parametric Kruskal-Wallis test to analyze the RES data.

4. Results

4.1. *Repeated Measures*

A repeated measures MANOVA showed a significant main effect of time with a large effect size, $F(2,136) = 82.23$, $MSE = 4.13$, $p < .001$, $r = .47$. Pairwise comparisons showed that there is a significant difference between the pre-test and the post-test ($p < .001$), between the pre-test and the follow-up test ($p < .001$), but not between the post-test and the follow-up test ($p = .14$). No main effects of condition on the tests can be found. The interaction between time and condition, however, was significant, with a small effect size, $F(4,136) = 5.63$, $MSE = 4.13$, $p < 001$, $r = .17$. Analyses of variance on the gain scores were conducted to follow up on the significant interaction.

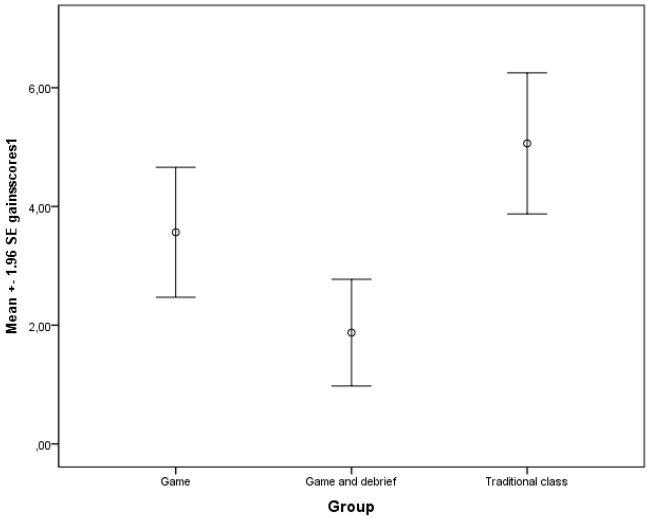
Figure 3: Line plot of scores on pre-, post and follow-up test.



4.2. *English vocabulary test: post-test 1*

Every condition showed a significant gain when comparing the first post-test to the pre-test ($p < .001$), showing that every condition entails a learning effect. An ANOVA on the gain scores revealed a significant effect of condition on the gain scores with a moderate to large effect size, $F(2,68) = 8.69$, $MSE = 7.02$, $p < .001$, $r = 0.45$.

Figure 4: Error bar graph of mean gain scores at post-test 1.



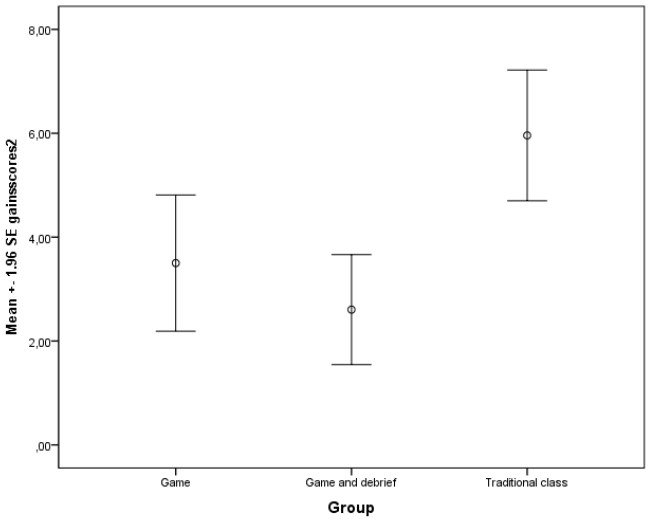
The highest gain in scores can be found in the traditional classroom condition ($M = 5.06$, $SD = 2.97$), followed by the game condition ($M = 3.57$, $SD = 2.68$). The game + debriefing condition showed the lowest gain in scores ($M = 1.88$, $SD = 2.25$). The Scheffé post-hoc test showed that only the differences between the traditional classroom condition and the game +

debriefing condition are significant ($p < .001$). The differences in gain scores between the traditional classroom and game condition were not significant ($p = .16$), nor between the game and game + debriefing condition ($p = .10$). H1 can thus be accepted as no significant differences can be observed between the group that only played the game and the traditional classroom condition. H1a has to be rejected, as a debriefing session did not add value to progress on the vocabulary test in comparison to participants who only played the game.

4.3. *English vocabulary test: post-test 2*

Every condition showed a significant gain when comparing pre-test to follow-up test ($p < .001$). An ANOVA on the follow-up tests showed statistically significant differences between the groups with a moderate to large effect size, $F(2,68) = 7.98$, $MSE = 9.06$, $p < .001$, $r = 0.44$).

Figure 5: Error bar graph of mean gain scores at post-test 2.



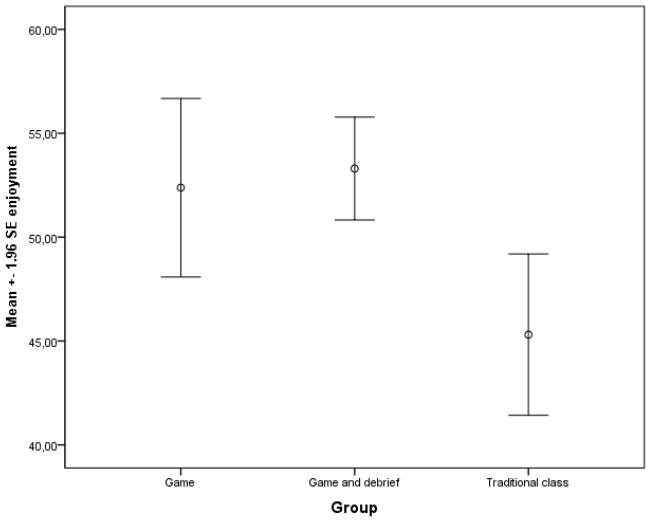
Again, the traditional classroom group showed the highest gain in scores from pre- to second post-test ($M = 6.00$, $SD = 3.15$), followed by the game group ($M = 3.5$, $SD = 3.21$). The game + debriefing condition showed the lowest gain in scores ($M = 2.6$, $SD = 2.65$). A Scheffé post-hoc test showed that the traditional classroom condition has significantly higher gain scores on the second post-test compared to both the game group ($p = .03$) and the game + debriefing group ($p < .001$). The difference in gain scores between the game and game + debriefing group was not statistically significant ($p = .60$). Hence, H3 is rejected as we expected the participants that only played the game, would have a similar gain at follow-up as participants who received the traditional class. Moreover, H3a is rejected, as we expected that the game +

debriefing would outperform the participants who only played the game and those in the traditional class.

4.4.*Enjoyment*

A Kruskall-Wallis test on the enjoyment scores showed a statistically significant difference between conditions with a small effect size, H(2) = 9.29, p = .01, r = .11. The game + debriefing group scores were the highest with a mean rank of 42.20, followed by the game group with a mean rank of 39.41. The traditional class scores were the lowest on the relative enjoyment scale with a mean rank of 25.33. Pairwise comparisons showed that the game + debriefing group scored significantly higher on the RES than the group receiving the traditional class (p = 0.01). A marginal significant difference was detected between the traditional class and the game group (p = .053). No differences were found between the game and game + debrief condition. We can thus accept H2. We, however, have to reject H2a, as a debriefing did not add value to enjoyment scores compared to participants who only played the game.

Figure 6: Error bar graph of mean scores on Relative Enjoyment Scale.



5. Discussion

The present study aimed at investigating the longer term effectiveness of DGBL and finding empirical evidence for the added value of a debriefing session in a DGBL context using an experimental approach. In the standardized procedure developed by the authors, the importance of controlling for time exposed to content and the content itself is stressed in order to maximize internal validity, which is something that is often neglected in published research (All, Nuñez Castellar & Van Looy, 2014; Randel et al., 1992; Sitzmann, 2011). Accordingly, in the present

study, we strictly controlled that during the three interventions children were instructed exactly the same content and spent an equal amount of time with the intervention.

The major contribution of this paper are the results of the follow-up study. While the DGBL platform participants showed similar learning gains at post-test 1, the traditional class outperformed them at post-test 2. Consequently, it cannot be considered as effective as a traditional class regarding learning outcomes as we believe that the follow-up test gives us a better indication of actual learning. This is in line with Clark (1985) who concluded the strongest effects for technology enhanced instruction are found in short-term studies, where post-tests are introduced shortly after the instruction. Our results are, however, not aligned with previous results, where non-significant differences between DGBL and a traditional lecture, become significant in favor of DGBL at follow-up (Brom, et al., 2011). However, the design by Brom and colleagues (2011) was different compared to our study in the sense that all children in the study (i.e., both game and traditional class condition) received an expository lesson before playing the game (game condition) or receiving an additional class (traditional class). The aim of both interventions was thus to reinforce the knowledge acquired during the expository lecture. This definitely brings forward the need for follow-up studies in DGBL effectiveness research, even if it is only after 3 weeks.

We assumed that a DGBL intervention is considered effective if it scores equally good on learning outcomes compared to a traditional method without significantly diminishing motivational or efficiency outcomes (All, Nuñez Castellar & Van Looy, 2015). Based on the results of the present study, -name game removed for blind review purpose- can be partly considered effective, considering that it scores higher on enjoyment. However, we need to note that our chances for a Type I error are increased when accepting this hypothesis, as we have used single imputation using the estimation maximization algorithm on 3 of the 12 items of the RES were missing values were up to 9,9%, resulting in lower standard errors and consequently, lower p-values. When DGBL is more fun and consequently, more motivating to play, this is a decisive factor for adoption/implementation, but only if similar learning gains are achieved compared to the traditional method, which is not the case here (All, Nuñez Castellar & Van Looy, 2015). Higher levels of motivation have also been brought forward as a required finding in order to justify the investment of developing and implementing DGBL, but again when no differences with regard to learning outcomes can be found (Clark, 2007). Considering that the timespan of the implemented interventions was short and limited (one thematic session of ten), we cannot make claims on whether or not it is justified to invest in the implementation of the DGBL platform in a class context. The reason for this is that the higher enjoyment scores for

the game groups, imply that the desire to continue playing will also be higher in this group (Schønau-Fog & Bjørner, 2012). Hence, it could be that if the DGBL platform is implemented over a longer period of time or could be played freely afterwards, this would result in additional exposure to the content in the game group and thus additional learning gains. Also, the platform is already commercially available and considering that English is not part of the curriculum in primary school, we have no materials we can compare it to. Considering an equal amount of time was spent on both the game and traditional intervention, efficiency outcomes regarding time management can be considered as equal, which is an important desired outcome of DGBL in a school context (All, Nuñez Castellar & Van Looy, 2015). However, as this was an experimental set-up, the teacher did not have to expense any effort in preparing the DGBL so our situation was not representative in terms of time-management as an efficiency outcome.

While the importance of a debriefing session is stressed in DGBL literature (Crookall, 2014; Garris, et al., 2002; Van Der Meij, et al., 2013) in order to enhance learning outcomes and stimulate a positive learning experience, the results of the present study do not support these statements. A debriefing session did -in this case- not add value to the game, neither for performance on the shorter or longer term, nor for enjoyment. We should, however, note that considering –name game removed- is an extrinsic game type, learning content is rather simple (memorizing words) and made explicit in the game-based activities, making the connection between game content and classroom content -which is the aim of a debriefing session- possibly unnecessary. Hence, further research into the added value of debriefing in a DGBL context is required, considering that being a form of experiential learning might not be a sufficient reason to implement a debriefing session in a DGBL context. More specifically, nuances regarding game type (intrinsic vs. extrinsic), complexity of learning content, explicit/implicit learning goals, game characteristics and possibly other factors need to be explored. Another alternative explanation is that the debriefing led to less in-game exercise time distracting them from the content, focusing on their own experiences on the expense of study time. However, the pupils in both game conditions are exposed to all the words in the first activity. The only aspect that differed in the game and game and debriefing condition, is the amount of repetition of that vocabulary. Hence, we believe our results are still valuable considering every condition was exposed to and rehearsed the vocabulary in the same timeframe allowing us to get an indication on which was most effective. On the other hand, our results are in line with the study conducted by Barzilai and colleagues (2014), where the group receiving a debriefing scored equally good on a knowledge test in comparison to participants who only played the game and had similar scores regarding enjoyment. Remarkably, in Barzilai's study, the participants in the debriefing

condition even spent additional time with the learning content (the debriefing was supplemented to gameplay time) and it still did not add value. Hence, more empirical studies investigating the added value of a debriefing session in a DGBL context are required.

In conclusion, we can draw three lessons from this validation study for the further development of our standardized procedure. Firstly, a follow-up test is indispensable for conducting effectiveness research on DGBL in order to determine whether initial effects can be sustained over time. Secondly, subjects' gameplay should be logged in order to get a better indication of which content they have been exposed to or to check whether or not differences in gameplay can be found between groups, in order to establish why certain conditions score better or worse. Finally, we have established that adding extra elements during the implementation of DGBL, such as a debriefing session, can yield different results and makes it impossible for researchers to get an indication of the educational game as such, without the addition of the extra elements. Consequently, in the context of DGBL effectiveness research, DGBL should be implemented as a stand-alone intervention in order to establish the effectiveness of the game as such. If one does add elements to the intervention, effectiveness claims can only be made regarding the intervention as a whole.

6. Limitations and further research

A first limitation of this study is that our sample is not large enough to be representative for the population of primary school children from the fourth and fifth grade. The timespan of the implemented interventions was also limited (one thematic session out of ten). Hence, more empirical studies with larger sample sizes and longer interventions are required in order to validate our results. A second limitation is that the study was conducted among pupils from the same school in Flanders. Results can thus be biased due to school influence. Consequently, an interesting venue for further research would be whether similar results are achieved for pupils from other schools and in other countries.

A third limitation in our study is that we do not have any indication with regard to the in-game progress of the participants that used the DGBL platform. So it could be, that participants in the game + debrief group did not complete the six activities. This can be a potential confound in our study. Another potential confound in our study is that in the traditional classroom group only the teacher was present, in the game + debrief group the teacher, ICT coordinator and researcher and in the game group the ICT coordinator and researcher. Hence, it could be that

pupils in the traditional classroom group were more at ease, while in the other groups the ‘out of the usual’ composition or presence of the researcher might have impacted results.

Lastly, our debriefing results cannot be generalized, because 1) we used an extrinsic game type 2) it only focused on a simple task (memorizing words) and 3) learning content was made explicit in the DGBL platform. Hence, further research testing the added value of a debriefing session using intrinsic game types (i.e., learning contents are integrated in the narrative or gaming mechanics) aimed at teaching more complex learning content (i.e., on a more conceptual level) and where implicit learning takes place, for which debriefing sessions are more suitable and are expected to impact effectiveness in a positive way. Another interesting venue for further research is to investigate implementation methods of DGBL, empirically testing the added value of adding elements, such as in this case, adding a debriefing session. This may especially be interesting when non-significant results are found when DGBL is implemented as a stand-alone method. This way, recommendations with regard to implementation of DGBL could be provided.

CHAPTER 6.
Pre-test influences on the effectiveness of digital-game based learning: a case study of a fire safety game

Abstract

In recent years, critiques have been formulated regarding current evaluation methods of digital game-based learning (DGBL) effectiveness, raising doubt with regard to the validity of certain results. A major issue of contention is whether or not a pre-test should be administered, gauging for baseline measures of knowledge that are targeted using an educational intervention. A pre-test session is considered beneficial as it allows researchers to control for pre-existing differences between the experimental and control group and facilitates the comparison of progress (i.e., gain scores) as a result of the intervention implemented. However, adding a pre-test can also result in a 'practice effect', meaning that subjects who take the same test twice, automatically perform better the second time. Moreover, pre-test sensitization can occur, which refers to the subjects being more sensitive to an intervention as a result of the pre-test. Hence, in a standard experimental setup one cannot know whether a positive effect as a result of the intervention would have been present if a pre-test had not been administered.

The present study aims to explore the advantages and disadvantages of adding a pre-test in DGBL effectiveness research. For this purpose, an effectiveness study of a fire safety training in a hospital was conducted using a Solomon four-group design. The experimental groups received a game-based intervention (N = 66) of which one group received a pre- and a post-test (n=25) and one group only received a post-test (n =41). The control groups received traditional classroom instruction (n=58), of which one group received a pre-and a post-test (n=39) and one group only received a post-test (n=29). A 2x2 ANOVA was used to explore the testing effect and the interaction between the pre-test and the intervention. No main effect of testing was found. However, an interaction effect between pre-test and intervention was detected. More specifically, this interaction takes place in the traditional classroom group: subjects who have received a pre-test in this group score significantly higher ($p < .05$) on the post-test than subjects in the traditional classroom group who did not receive a pre-test. This was not the case in the game group.

When the administration of a pre-test influences the control group's receptivity to the intervention, but not that of the experimental group, results of an effectiveness study may be biased. Hence comparison of post-test scores of different treatments in pre-test/post-test designs may be problematic. This is an important finding in the context of DGBL effectiveness research as the presence of a pre-test may artificially inflate the learning outcomes of the control condition. Therefore, further research should take this into account and look for possible solutions to solve this discrepancy. However, in the present study, we were able to show that the game was highly effective, as both game groups still outperformed the slide-based group that received a pre-test. The Solomon four group design has thus shown its added value and more effectiveness studies on DGBL implementing this design are required in order to further validate these results.

Keywords:

Digital game-based learning; effectiveness assessment; Solomon four-group design; practice effect; pre-test sensitization

Reference:

All, A., Plovie, B., Nuñez Castellar & Van Looy, J. (In review) Pre-test influences on the effectiveness of digital-game based learning: a case study of a fire safety game.

Preliminary results presented at the 66th annual conference of the International Communication Association 'Communicating with power', Fukuoka, Japan: All. A, Plovie, B., Nuñez Castellar, E.P. & Van Looy, J. (2016). *Testing the effects of administering a pretest on the effectiveness assessment of a hospital fire safety game*

1. Introduction

The interest in using digital games as instructional tools has increased strongly over the past decade. Digital game-based learning (DGBL) has been implemented in various sectors such as defense, education, corporate training, health and wellbeing, and communication (Backlund & Hendrix, 2013). Concomitantly there has been a growing interest and production in research into DGBL's effectiveness (Mayer et al., 2014; Wouters, Van Nimwegen, Van Oostendorp, & Van Der Spek, 2013). Results have failed to provide conclusive evidence however (Backlund & Hendrix, 2013; D. B. Clark, Tanner-Smith, & Killingsworth, 2015; Wouters et al., 2013). This is at least in part due to the variety of shapes that DGBL presents in terms of topics and game genres (Kirriemuir & McFarlane, 2004). Another important factor has been the heterogeneity in study designs for assessing their effectiveness however (All, Nuñez Castellar & Van Looy, 2014). Research designs differ in several respects including use of a control group, activities presented in the control group(s), implementation of DGBL (stand-alone vs. in a broader program), outcome measures to assess effectiveness, statistical techniques to quantify learning outcomes and the administration of a pre-test (Girard, Ecalle, & Magnan, 2013, All, Nuñez Castellar & Van Looy, 2014). Moreover, methodological issues have been brought forward regarding published effectiveness research on DGBL (Clark, 2007; D. B. Clark et al., 2015; Girard et al., 2013, All, Nuñez Castellar & Van Looy, 2014), which sometimes lacks rigorous assessment (Clark, 2007; D. B. Clark et al., 2015; Connolly, 2014). For instance, studies are frequently being implemented without a strict control of potential threats to their internal validity, such as the addition of training materials to the intervention (e.g., required reading, exercises) or the lack of a standardized protocol for instructors (e.g., procedural help, guidance only during the intervention). Moreover, authors regularly fail to mention whether or not self-developed tests have been piloted, which leads to uncertainty with regard to the reliability and validity of results (Brom, Šisler, Buchtová, Klement, & Levčík, 2012). Another important methodological issue is that it is difficult to replicate published DGBL effectiveness studies given that authors often do not provide sufficient information on how the intervention – both in the experimental and control condition – has been implemented (Sitzmann, 2011; All, Nuñez Castellar & Van Looy, 2014). Detailed information on procedure is indispensable, however, in order to gain insight in whether the gains that are reported are a consequence of the different methods and not due to other circumstantial factors that differed between conditions

(Randel, Morris, Wetzel, & Whitehill, 1992) and to be able to replicate studies (All, Nuñez Castellar & Van Looy, 2014).

Considering these methodological limitations, a more systematic approach that can serve as a guideline for quality assessment is required for researchers willing to conduct effectiveness studies in this field (Mayer et al. 2014). For this purpose, research into preferred study designs is required. In the present study, we aim to investigate whether or not a pre-test of knowledge should be administered, as the absence of a pre-test is one of the main criticisms of DGBL effectiveness studies (Clark, 2007; O'Neil, Wainess, & Baker, 2005; All, Nuñez Castellar & Van Looy, 2016) and studies without a pre-test are consequently often omitted from meta-analyses on DGBL effectiveness (e.g., D. B. Clark et al., 2015; Girard et al., 2013)

1.1. *Pre-test administration*

Administration of a pre-test is contentious topic as gauging for baseline measures of knowledge can provide additional data regarding participants but it can also threaten the validity of results. Adding a pre-test to the research design is useful as it allows researchers to control for pre-existing differences between the experimental and control group (Clark 2007) and to compare progress (i.e., gain scores) as a result of the interventions (Gerber and Green 2012). By adding pre-test scores to the analysis – for example when comparing gain scores or conducting repeated measures or analysis of covariance with pre-test scores as covariate – error variance is also reduced, resulting in a more precise estimate of the treatment effect, allowing for the use of statistically more powerful tests (Dimitrov & Rumrill, 2003; Knapp and Schafer 2009). Lastly, the addition of a pre-test allows the researcher to control for characteristics of drop-outs (All, Nuñez Castellar & Van Looy, 2016) so that potential biases with regard to representativeness of the sample can be reported. On the other hand, adding a pre-test can also 'blur' the real effect of the treatment. Firstly, administering a pre-test can result in 'practice effects', meaning that subjects who take the same test twice, may do better the second time, even if the intervention had not taken place (Crawford et al. 1989). In this case the effect is due to the pre-test, as it can offer participants additional exercise materials or item training (van Engelenburg 1999). Hence progress due to the intervention and progress due to the practice effect cannot be isolated from each other. Moreover, pre-test sensitization can occur, referring to an interaction effect of the pre-test and the treatment (Braver and Braver 1988; van Engelenburg 1999). This means that subjects who have received a pre-test will be more sensitive to the intervention as compared to subjects who have not received a pre-test, resulting

in higher scores on the post-test. For instance, when implementing the same test pre- and post-intervention during a short period of time, a pre-test can cue students on what should be remembered from the intervention (Randel, Morris, Wetzel, & Whitehill, 1992). Consequently, one cannot know whether a positive effect as a result of the treatment would have been present if a pre-test had not been administered. In that case generalization of results from a pre-tested to an un-pretested sample is made impossible. This has resulted in researchers renouncing a pre-test when studying effectiveness of DGBL (e.g., Amory 2010; Tsai et al. 2012). However, to our knowledge, pre-test influences have hitherto never been studied in a DGBL context. Therefore, before making assumptions on the presence of a practice effect or pretest sensitization, this needs to be studied (Braver and Braver 1988).

An experimental design that is proposed to investigate the issues of practice effects and pre-test sensitization is the Solomon four-group design (Solomon 1949). Thereby four conditions are proposed: the first two conditions are the same as in the classic pretest-posttest control group design: participants receive a pretest, an intervention (experimental or control) and a post-test. The two extra conditions duplicate the treatment and control conditions, but a pre-test is absent. The present study aims to assess pre-test influences in a DGBL effectiveness research context. More specifically, we aim at testing for a main effect of pre-test (i.e., pre-test effect) and an interaction effect between pre-test and treatment (i.e., pre-test sensitization) on learning outcomes.

2. Method

2.1. *Design*

A Solomon four-group design was implemented in order to assess the effectiveness of a digital game-based fire safety training among hospital personnel. Participants in the experimental condition received a digital game-based intervention and participants in the control group received the traditional slide-based lecture. Individual randomization of subjects was not possible in this study due to practical limitations, as staff needs to subscribe for the fire safety training (i.e., the hospital disposes of a large pool of staff such as nurses, cleaning personnel, doctors, etc. who work in shifts). Hence, randomization was implemented on a group level (i.e., a group was composed of people that subscribed for a safety training on the same date).

2.2. *Stimulus material*

2.2.1. Digital game-based fire safety training

The DGBL fire safety training was specially developed for the hospital whose personnel participated in the study. All hospital personnel (i.e., doctors, nurses, cleaning personnel, administrative staff, technical staff, etc.) is required to complete the fire safety training every year. Because the hospital has expanded over the years and is still expanding and personnel works in different shifts, it is becoming increasingly difficult to organize traditional training for everyone. Hence the decision to develop a digital game in cooperation with DAE Research (HOWEST). The game consists of three mini-games or courses: 'small fire'; 'smoke' and 'blaze'. After participants have completed these courses, they can also play a random 'fire safety' scenario, during which elements learned in the course can be practiced. In total, 6 different scenarios are available. The game can be freely played on the following website: http://sggo.howest.be/het-serious-game.

2.2.2. PowerPoint

The PowerPoint lecture is instructed by the prevention manager of the hospital. This is the lecture that is currently being used as a fire safety training for the hospital personnel. This lecture was also used as a basis to define content treated in the game and contains exactly the same material as treated in the game.

2.3.*Procedure*

2.3.1. Experimental groups

The experimental groups played the game in a conference room on one of the four campuses of the hospital during working hours. A maximum of six subjects could participate per session. When entering the conference room, subjects received an introduction by a researcher with information by the researcher regarding the purpose of the study. Afterwards, the subjects either filled out the pre-test (experimental condition with pre-test) or started playing the game (experimental condition without pre-test). The subjects played the game individually on a laptop computer with a headphone. During game play, two researchers were present providing procedural help, meaning that only technically oriented help was provided when there were issues with the computer or game play (i.e., no help regarding course materials). After the subjects completed all three courses and one scenario, a post-test was administered. In total, 18 game training sessions were organized; 9 included a pre-test and 9 did not.

2.3.2. Control groups

The control groups equally received the slide-based lecture in a conference room on one of the four campuses. The slide-based lecture was given by either the prevention manager or another fixed employee from the prevention staff that was responsible for the fire safety training. The same procedures were followed regarding administration of pre-test and post-test as in the experimental groups. The subjects were instructed in groups of minimum 8 and maximum 20 people. In total, 6 slide-based lectures were organized, 3 included a pre-test and 3 did not. During every slide-based lecture, the same two researchers who were present during the DGBL intervention were present during the slide-based lectures to check whether all topics discussed in the game were also discussed in the slide-based lecture using a topic list (see appendix A).

2.4.*Participants*

The present study was conducted in collaboration with the hospital AZ groeninge in Kortrijk, Belgium. In total, 152 subjects participated in the study. Eighty-three subjects participated in the experimental groups, of whom 42 received a pre-test and 41 did not receive a pre-test. Sixty-nine subjects participated in the control groups of whom 39 received a pre-test and 30 did not receive a pre-test. Nineteen subjects in the experimental group (8 who received a pre-test and 11 who did not receive a pre-test) were excluded from the analysis because log data showed that they either did not complete all three courses or they repeated a course several times. In the end, 133 participants were retained for the analysis.

As can be seen in table 1, randomization on a group level has led to a balanced group in terms of age and proportion of gamers, but not in terms of gender composition.

Table 1: Control for balanced groups as a result of randomization on group level

	Experimental group with pre-test (n=34)	**Experimental group without pre-test (n=31)**	**Control group with pre-test (n=39)**	**Control group without pre-test (n=29)**	**Chi²/ F**	**p**
Women	76.50%	71.00%	92.30%	96.60%	10.87	**.01**
Age (mean)	40.03	37.52	38.31	40.83	.54	.66
Gamers	50.00%	61.30%	61.50%	48.10%	2.00	.57

2.5.*Measures*

Three types of outcomes should be considered when assessing effectiveness of DGBL: learning outcomes, motivational outcomes and efficiency outcomes (All, Nuñez Casteller &

Van Looy, 2015). DGBL is considered effective if it succeeds in achieving similar learning outcomes compared to more traditional methods, without significantly diminishing any of the others. In the present study, we have assessed performance as an indicator for cognitive learning outcomes, motivation towards the instructional material as an indicator for motivational outcomes and time exposed to intervention as an indicator for efficiency outcomes.

2.5.1. Cognitive learning outcomes

In order to assess performance, a test was developed by the researchers in cooperation with the prevention staff responsible for the fire safety training – the same staff who provided the slide-based lectures. The test had previously been implemented in a pilot study (N = 52) in an initial phase of development of the game. The test consisted of 18 open-ended questions, covering all topics that are treated in the interventions allowing for a maximum score of 40. The test assesses declarative and procedural knowledge. Examples of questions are*: What is the first step you have to take when a small fire breaks out? How do you do this? What are the three steps to follow when evacuating patients? Which three steps do you have to take to evacuate a bedridden patient? Etc.* The tests were corrected by two researchers. For this purpose, an evaluation form was developed in order to guarantee a standardized manner of correcting the tests. If there was uncertainty regarding the correctness of certain answers, the correctors discussed the response and agreed upon a score.

2.5.2. Motivational outcomes

The Instructional Materials Motivation Survey (IMMS, Keller 1987) was used to assess motivation towards the instruction method. We based ourselves on Huang and colleagues (2010) for the game version of the IMMS. The IMMS consists of 36 items, divided in 4 subscales: attention (i.e., gaining and keeping the learner's attention), relevance (i.e., activities must relate to current situation or to them personally), confidence/challenge (i.e., activities cannot be perceived as too hard or too easy, which is also a prerequisite for an optimal game experience or game flow) and satisfaction/success (i.e., learners must attain some type of satisfaction or reward from the learning experience). The items were scored on a 5-point Likert scale, with 1 being 'not true' to 5 'very true'. The total score represents motivation towards the instructional material. The scores on the subscales give an indication as to the sub dimensions on which the intervention was either more or less successful (Keller 2010).

A reliability analysis of our data showed an acceptable Cronbach's alpha for subscales attention ($\alpha = .82$) and satisfaction ($\alpha = .84$), but not for subscales confidence ($\alpha = .50$) and relevance (α

= .67). Hence, we deleted confidence items 1 (‘When I first looked at the game/slides, I had the impression that it would be easy for me’) and 34 (‘I could not really understand quite a bit of the material in the game/slides’.) We have also deleted relevance item 26 (‘the game/lecture was not relevant to my needs because I already knew most of it’). With these items deleted, confidence (α = 68) and relevance (α = .70) have an acceptable Cronbach’s alpha.

2.5.3. Efficiency outcomes

Time management as an efficiency outcome refers to DGBL succeeding in reducing the timeframe needed to teach a certain content matter (All, Nuñez Castellar & Van Looy, 2015). Hence, we have timed every separate slide-based lecture and retrieved individual information on total time spent on the DGBL intervention based on automated logging.

2.6.*Data analysis*

Assumptions for normality were checked by comparing the numerical value for skewness and kurtosis with the respective standard error (Field, 2009) and checking the Q-Q plot of standardized residuals of the dependent variable (Kutner, Nachtsheim, Neter, & Li, 2005). Since the post-test scores were not normally distributed (i.e., negatively skewed), a reversed square root transformation was applied on the post-test data for the analyses of variance (Field, 2009). The difference scores were normally distributed, so no transformation was necessary for the paired samples t-test (Field, 2009). Effect size r was used which for t-test analyses was calculated by taking the square root of $t^2/(t^2+df)$ (Field, 2009). For analyses of variance effect size was calculated by taking the square root of eta squared which was obtained using the following formula: $SS_{between}/SS_{total}$ (Field, 2009; Levine & Hullett, 2002). To check for equality of variance, Levene’s test was used (Kutner et al., 2005).

3. Results

Firstly, we will discuss the effectiveness of the DGBL intervention and secondly, we will discuss the influence of the pre-test on outcome results.

3.1. *Effectiveness of the DGBL treatment*

Considering that we can distinguish two designs (pre-test post-test design and post-only) in our data (van Engelenburg, 1999) in order to assess the effectiveness of DGBL, we will conduct analyses on two datasets: one containing the participants receiving both a pre- and a post-test and one on the participants only receiving a post-test.

3.1.1. Pre-test post-test design

A paired samples t-test (N = 73) showed a difference between pre- and post-test scores for both the participants receiving a slide-based lecture, $t(38) = 9.45$, $p < .01$, $r = .84$, and the participants receiving the DGBL intervention $t(33) = 12.78$, $p < .01$, r = .91, showing that both produce a large learning effect. Table 2 provides an overview of the descriptive statistics of the pre- and post-test scores of both instruction groups.

Table 2: Descriptive statistics of pre-test post-test design (N = 73)

Group	***N***	***Pre-test score (M/SD)***	***Post-test score (M/SD)***	***Minimum score (pre-test/post-test)***	***Maximum score (pre-test/post-test)***	***Adjusted post-test score (M/SD)***	***Gain score (M/SD)***	***Time spent on the intervention***	***Total score IMMS (M/SD)***
Game group	34	16.19/7.8	34.44/5.64	3.5/17	33/39.5	33.75/.84	18.03/5.45	35,08 min	4.21/.48
Lecture group	39	9.27/5.2	27.29/5.31	1/18	14.5/35	28.63/.91	18.25/7.30	25,18 min	3.86/.48

In order to compare the effectiveness of the DGBL treatment, we first checked for pre-existing differences by conducting an analysis of variance (ANOVA) with pre-test as dependent and instruction method as independent variable. Results show that the DGBL group scored significantly higher on the pre-test than the group that received a slide-based lecture, $F(1,71) = 20.31$, $p < .01$. Two types of analyses can be distinguished in the literature which allow to take these pre-existing differences into account: an ANOVA on the change scores and an analysis of covariance (ANCOVA) with pre-test scores as covariate (Dimitrov & Rumrill, 2003; Knapp & Schafer, 2009). Considering that there is no agreement on which one to use and that the aim

of the present study is to explore the disadvantages and advantages of adding a pre-test in your study design, we provide results for both.

Results of the ANCOVA show that, after controlling for initial differences on the knowledge test, instruction type still has an effect on the post-test scores showing a medium effect size (1,71) = 18,357, $p < .01$, $r = .35$. More specifically, the group receiving the DGBL treatment outperformed participants receiving the slide-based lecture on fire safety knowledge (see table 2). Consequently, based on the ANCOVA, we can state that the DGBL treatment is more effective for the fire safety training than the slide-based lecture regarding learning outcomes. When we conduct an ANOVA on the change scores, however, we do not find a difference $F(1,71) = .22$, $p = .88$, $r = .02$. This would imply that both groups showed a similar learning gain as a result of the intervention (see table 2).

An ANOVA on the post-test scores of the Instructional Materials Motivation Survey shows a significantly higher score for the groups who received the game-based intervention with a medium effect size, $F(1,66) = 8.64$, $p = .01$, $r = .34$. When looking at the subscales, a difference can be found for confidence ($p < .01$, $r = .37$), satisfaction/success ($p < .01$, $r = .35$) and for attention ($p = 0.07$, $r = .25$) but not for relevance ($p = .12$, $r = .19$). All participants thus perceived both the slide- and game-based method as relevant to their professional or personal context. However, the participants who received the slide-based lecture felt less satisfied with the learning experience (i.e., did not feel rewarded for it) and perceived more of an imbalance between knowledge/skills and the challenge that the instruction brought forward, compared to the game-based intervention. Moreover, participants in the DGBL conditions felt that the game succeeded more in gaining and keeping their attention during the intervention compared to participants in the lecture condition.

Regarding time management, an ANOVA on the time spent shows that the participants receiving a slide-based lecture spent a significantly higher amount of time on the intervention, with a large effect size $F(1,71) = 54,61$, $p < .001$, $r = .66$. More specifically, the lecture took on average 9.17 minutes longer. Consequently, the game-based intervention is more effective regarding efficiency outcomes.

3.1.2. Post-only design (N = 60)

When we compare the post-test data of participants who only received a post-test, an effect of treatment on performance can be detected in favor of the DGBL intervention with a large effect size, $F(1,58) = 104,233$, $p < .001$, $r = .80$.

No difference can be found for the IMMS, $F(1,54) = 2,606$, $p = .11$, $r = .2$. When we look at the subscales, however, a difference can be found for satisfaction in favor of the DGBL training, $F(1,56) = 5,021$, $p = .03$, $r = .29$.

Regarding time management, an ANOVA on the time spent shows that the participants receiving a slide-based lecture spent a significant amount of time more on the intervention, with a large effect size $F(1,55) = 46,42$, $p < .001$, $r = .68$. More specifically, the lecture took on average 10 minutes longer. Consequently, the game-based intervention is more effective regarding efficiency outcomes.

Table 3: Descriptive statistics of the post-test only design

Group	***N***	***Post-test score (M/SD)***	***Minimum score post-test***	***Maximum score Post-test***	***Time spent on the intervention***	***Total score on IMMS (M/SD)***
Game group	31	35.05/3.78	25.50	39.5	35 min	4.14/.36
Lecture group	29	21.60/6.65	7	31.5	24,59 min	3.98/.41

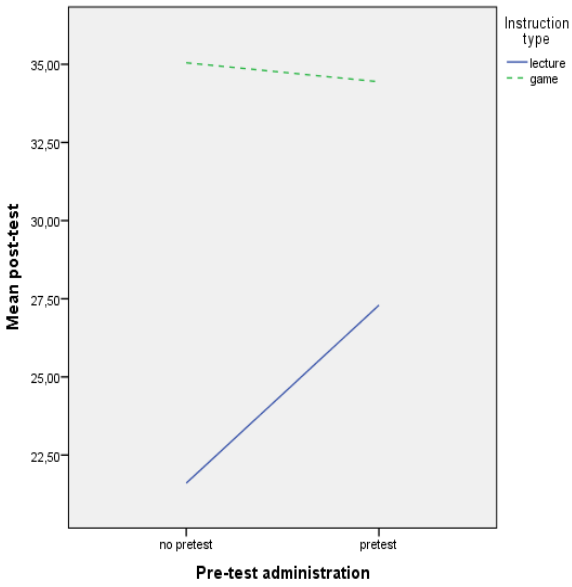
3.2. *Effect of the pre-test*

Considering that there are no guidelines for the types of analyses to conduct when pre-existing differences exist in a Solomon 4-group design, we conducted our analysis twice: once with the complete data set (i.e., including individual differences on the pre-test of knowledge) and once matching participants who received a pre-test on their pre-test scores ($N = 102$).

3.2.1. Analysis on complete dataset

In order to assess the influence of the pre-test on both the post- test and the treatment, we conducted a 2x2 ANOVA as suggested by Braver & Braver (1988). The two independent factors where the administration of a pre-test (two levels: pre-test was administered or no pre-test was administered) and the instruction type (two levels: DGBL or slide-based lecture). The dependent variable was post-test score. All statistics below are based on the transformed data, but the graphs reflect the untransformed data. Results show that there is a very large main effect of instruction type $F(1,129) = 136.67$, $p < .01$, $r = .71$. More specifically, the participants that received the DGBL intervention scored significantly higher on the post-test. The results show a small main effect of administering a pre-test $F(1,129) = 5.22$, $p = .02$, $r = .14$ and an interaction between pre-test and instruction type with a small effect size $F(1,129) = 7.46$, $p = .01$, $r = .17$. In fig. 2, we see that the influence of the pre-test on the treatment is larger in the group that received a slide-based lecture than in the group that received a DGBL intervention.

Fig. 1: Line plot of mean post-test scores (N=133)



When we compare post-test scores of the four groups using an ANOVA with the grouping variable (four levels: DGBL with pre-test, DGBL without pre-test, slide-based with pre-test and lecture without pre-test) again we see a very large effect of instruction method, $F(3,129) = 47.44$, $r = .72$. A post-hoc Scheffé test shows that no differences can be found between the DGBL group that received a pre-test and the DGBL group that did not receive a pre-test ($p = .99$). A difference is detected between the slide-based lecture group that did receive a pre-test and that which did not however ($p < .01$). More specifically, the group that received a pre-test before receiving the slide-based lecture, scores significantly higher than the group that did not receive a pre-test before the slide-based lecture. This indicates that administering a pre-test influences the participants' sensitivity to receiving the fire safety training with a lecture, resulting in higher scores on the post-test. This is not the case when receiving the training by DGBL.

Both gaming groups still score significantly higher on the post-test scores compared to the lecture groups, indicating that the game is more effective in terms of knowledge transfer than the slide-based lecture.

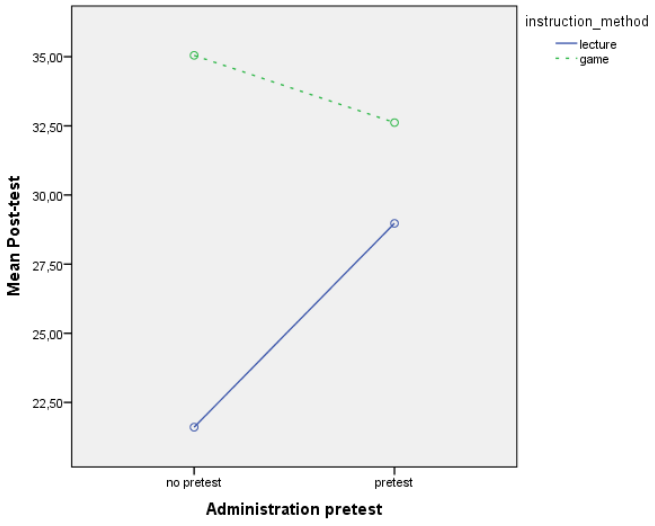
3.2.2. Analysis on matched groups

Matched groups were constructed for the participants that received a pre-test, by looking for participants in the DGBL and the slide-based lecture group that have a similar score (i.e., maximum 1 point difference) (Rubin, 1973). In the end, 21 participants remained in both the DGBL and lecture group that received a pre-test. No differences were found on pre-test scores between the newly composed experimental and control groups receiving a pre-test, $F(1,40) = .02$, $p = .82$. Since the other groups did not receive a pre-test, we could not match them based on pre-test scores and thus left them unmodified. The present analysis was conducted on a sample of 102 participants.

In order to test for pre-test influences, we conducted the same 2X2 ANOVA as discussed in 3.2.1. In line with the results on the complete dataset, we find a main large effect of instruction type, in favor of the DGBL intervention, $F(1,98)=46,29$, $p < .01$, $r = .59$. The main effect of pre-test, however, disappears $F(1,98) = 2.7$, $p = .1$, $r = .12$. An interaction between instruction type and pre-test administration is still detected, showing a small to medium effect $F(1,98) =16.42$, $p < .01$, $r = .28$. When we look at the graph, we again see that the influence of the pre-test on the treatment is larger in the group that received a slide-based lecture than in the group that received a DGBL intervention.

When conducting an ANOVA on the post-test scores of the 4 groups, instruction has a medium to large effect on post-test scores, $F(3,98) = 34.98$, $p < .01$, $r = .46$. A post-hoc Scheffé test shows that differences can be found between lecture with pre-test and the lecture without pre-test groups ($p < .01$). More specifically, the group that received a pre-test before the slide-based lecture scored significantly higher on the post-test than participants who did not receive a

Fig. 2: Line plot of mean post-test scores (N=102)



pre-test before the lecture. No differences are found between the game group that received a pre-test and the game group that did not receive a pre-test ($p = .41$).
The game groups significantly outperformed both lecture groups ($p < .05$), indicating that the game is more effective in teaching the fire safety training to the hospital personnel than the slide-based lecture in terms of learning outcomes.

3.2.3. In-game behavior

During the experiment, log data regarding in game actions were collected from participants in the game-based intervention group. During game play, players were able to retrieve 'information cards' regarding correct procedures. The number of times players opened these cards was logged by the system. These information cards are also opened automatically at the beginning of each training. Total time spent on studying the information cards (both opened automatically and manually) was logged. Mistakes made during the game and scores were also logged. These log data provide us with an opportunity to check whether participants receiving a pre-test before the DGBL intervention behave differently from participants who did not and thus objectively test whether or not pre-test sensitization took place. As can be seen in table 4, no differences can be found regarding number of information cards that were consulted manually, total time spent on these information, cards, in-game scores or total time spent on the intervention.

Table 4: Comparison of log data from DGBL groups

Log data	Pre-test before DGBL training	No pre-test before DGBL training	p
Number of information cards consulted manually	.48	.32	.67
Total time spent on information cards	239.79 s	228.32 s	.6
In game score for 'small fire'	872.15	758.67	.11
In game score for 'Smoke'	5806.29	5376.12	.41
In game score for 'Blaze'	3822.15	3485.21	.29
Total time spent on the DGBL intervention	1814,95 s	2001,67 s	.52

4. Discussion & conclusion

The aim of this paper was twofold. On the one hand, we studied the effectiveness of a digital game-based fire safety training compared to the traditional lecture-based type. On the other hand, we assessed the impact of the administration of a pre-test on learning outcomes by means of a Solomon four group design. We will first discuss the results of the pre-test impact in order to be able to more accurately interpret the results of the effectiveness study.

Our results revealed that the pre-test influence on an educational intervention depends on the type of instruction that is administered. More specifically, pre-test sensitization was detected among the participants receiving the more traditional slide-based lecture, but not among those receiving the DGBL intervention. Participants receiving a pre-test before receiving the slide-based lecture were thus more sensitive to the intervention and consequently scored significantly higher on the post-test than participants that did not receive a pre-test before the slide-based lecture. Providing a pre-test to participants receiving a DGBL intervention did not result in higher scores on the post-test compared to participants that did not receive a pre-test before the DGBL fire safety training. When receptivity to an intervention is altered due to the pre-test in one group and not in the group to which it is compared to, bias is introduced in the design (McCambridge et al. 2011). This is an important implication for the DGBL research field, as effectiveness studies on DGBL often show non-significant differences compared to traditional instruction (Backlund & Hendrix, 2013). In pre-test post-test designs this can lead to issues related to internal validity as post-test scores in control groups receiving traditional instruction might be significantly elevated as a result of administration of the pre-test while the scores in the DGBL treatment represent less biased scores. This non-significant difference might have been significant in favor of DGBL when no pre-test sensitization would have occurred in the

traditional lecture. This makes comparison of post-test scores as a result of different instruction methods difficult in pre-test post-test designs.

The results of our effectiveness study reflect the issue discussed above. While we could find an effect of treatment in favor of the DGBL training when comparing post-test scores of participants receiving both a pre- and post-test and participants who only received a post-test (after controlling for initial differences), we could not find an effect of treatment when comparing progress of the participants receiving both a pre- and a post-test. Considering that, based on our results, the post-test scores of the participants in the lecture group are positively biased and that both game groups (those receiving a pre-test before the intervention and those not) still outperformed the lecture group receiving a pre-test, we conclude that the DGBL fire safety training is more effective regarding learning outcomes.

Regarding motivational outcomes, results were mixed. While the DGBL group scored better on the IMMS when comparing treatments of participants receiving both a pre- and post-test, this result was not present among participants only receiving a post-test. Results regarding time management in favor of DGBL could, however, be replicated among both the participants who received a pre- and post-test and among participants only receiving a post-test.

A possible explanation for why pre-test sensitization takes place in the lecture group, but not in the game group can be found in a combination of the motivation paradigm of entertainment education proposed by Ritterfeld and Weber (2006) and self-determination theory (Ryan & Deci, 2000). According to the motivation paradigm, DGBL is implemented to 'seduce' the learners by gameplay to allocate their attention to the learning content. Interactivity is one of the main characteristics of game-based learning (and e-learning in general) resulting in higher attention during the activity and consequently, deeper processing of the content (Ritterfeld, Weber, Fernandes, & Vorderer, 2004). According to self-determination theory, higher levels of autonomy related to regulation for motivation to engage in an educational intervention leads to better performance (Ryan & Deci, 2000). Applied to our case study, this means that the pre-test initially primed both participants receiving the game training and the slide-based lecture to be more attentive due to a change of external motivation with external regulation (i.e., following the training because they are obligated) to introjected regulation (i.e., following the training as a need to prove their ability), leading to higher engagement during the training and higher scores on the post-test (Ryan & Deci, 2000). This can explain the higher post-test scores among participants receiving a pre-test before the slide-based lecture, compared to those who did not receive a pre-test. The interactivity of the fire safety game, however, requires the participants to focus their attention to the content to complete the course, regardless of whether they received

a pre-test or not. This can explain the similar scores and similar in-game behavior among participants receiving a pre-test before the DGBL training and those who did not. Moreover, the interactivity of the game training possibly stimulated a shift towards a more intrinsic motivation to finish the training for both game groups (Clark, 2007), which can explain the higher scores of the DGBL groups compared to the slide-based groups.

This is supported by our data which show a significantly higher score on the attention subscale of the IMMS in favor of the DGBL training when comparing treatments among participants who received both a pre- and post-test. We could not replicate this finding when comparing treatments among participants who only received a post-test. However, the researchers who were present during all interventions, perceived participants in the lecture groups who did not receive a pre-test before the intervention as more noisy and less attentive. Hence, it could be that socially desirable answers are blurring our data. A motivational finding consistent over all designs, however, was the difference between the game groups and lecture groups on the satisfaction subscale in favor of the DGBL training. This again, can be an argument in favor of a shift to intrinsic motivation in the DGBL groups, as a rewarding experience is considered a key characteristic of games for stimulating intrinsic motivation (Hwa Hsu, Lee, & Wu, 2005). Hence, the added value of DGBL in the present study can be found in its ability to 1) attract and keep the attention of the learning during the whole course of the intervention as a result of its interactivity and 2) stimulate intrinsic motivation during the intervention.

With the present study, we have also established the advantages of adding a pre-test, indicating pre-existing differences between experimental and control group. Consequently, when looking into the effectiveness of the DGBL treatment, we could control for these initial differences by adding pre-test scores as a covariate and thus have a more precise estimate of our treatment effect (Dimitrov & Rumrill, 2003).

Bearing in mind the advantages and disadvantages of the administration of a pre-test as described above, we have several recommendations for researchers aiming to assess the effectiveness of DGBL. Firstly, pre-tests should be administered but time between pre- and post-test should be increased, minimizing the influence of the pre-test (Dochy et al. 1999). This also gives researchers the opportunity to match participants in experimental and control group, based on their pre-test scores (Gerber and Green 2012). Secondly, we recommend researchers to not only report on differences between groups regarding progress (i.e., gain scores), but also on post-test scores, in order to provide a more complete understanding of the data, as these can yield different results (Knapp and Schafer 2009).

On a final note, while the Solomon design seemed to be promising, there is a lack of clear guidelines on how to analyze its data. In the present study we bumped upon the issue of partly failed randomization and thus higher scores on the pre-test in the experimental group. While we aimed to solve this issue by conducting an analysis on matched groups regarding pre-test scores, we still do not have an indication as to what extent the groups that did not receive a pre-test had the same baseline knowledge regarding fire safety. The Solomon design seems to be a good design if one has perfect data, but in the social sciences, data is often not textbook perfect and researchers do not have the opportunity to gather over thousands of data points assuring successful randomization. Nevertheless, in the present study, the Solomon design has proven its added value, as it provides us with a more nuanced view of our data. For instance, we can make a more supported claim on the DGBL effectiveness regarding learning outcomes and know that we have to be careful with the interpretation of our results regarding motivational outcomes.

5. Limitations and Further research

Further research implementing the Solomon design is required, as there were pre-existing differences between the experimental and control group in the pre-tested groups. This keeps us in doubt about the similarity of the experimental and control groups in the un-pretested groups regarding prior knowledge on fire safety, possibly influencing our results. Hence, further validation of our results is required. A limitation of the present study is that the group dynamics in the experimental and control groups are not the same (individual vs. group). However, as the aim of the hospital is to replace the traditional slide based lecture by the fire safety training game, the comparison we made is meaningful.

In hindsight, it would also have been interesting to add observational data on the control groups in our study to get insight into whether participants receiving a pre-test before the slide-based lectures actually behave differently than those that did not receive a pretest before the lecture as an indicator for pre-test sensitization. The researcher present during the intervention had the impression that participants receiving a pre-test before the slide-based lecture were more attentive than participants in the second control group, but these are subjective perceptions carrying only limited validity.

Finally, a follow-up study to see longer term effects of both interventions and the longer term effects of the pre-test would be of added value for the present study and will be.conducted later this year.

CHAPTER 7.
Testing the effectiveness of digital game-based learning in a corporate context: comparison to a passive e-learning approach

Abstract

The investment in human resources by means of training programs is a key factor in creating competitive advantage for a commercial enterprise. For cost-efficiency reasons, e-learning programs are increasingly being implemented. These programs, however, are not always being used by employees. The present study aims to test whether digital-games based learning can offer a solution for the non-engagement and drop-out of employees in e-learning programs. More specifically, the present study investigated whether the interactivity of a game results in higher motivation to learn using the method, higher levels of enjoyment and better learning outcomes compared to a passive, instructional video. For this purpose, an experimental study was conducted among 64 employees working at a large bank, testing an e-learning training program (game or instructional video) aimed at teaching the bank's basic client-oriented principles in order to improve their loyalty to the bank. No differences regarding motivation, enjoyment or learning outcomes were found between participants receiving the game training and the instructional video. This shows that it might not always be required to –in a corporate context- invest in interactive content, considering it was not able to overcome the motivational issues related to more traditional e-learning approaches.

Keywords:

Digital game-based learning, effectiveness, corporate training, performance, motivation

Reference:

All, A., Nuñez Castellar, E.P. & Van Looy, J. (In review) . Testing the effectiveness of digital game-based learning in a corporate context: comparison to a passive e-learning approach

Preliminary results presented at the 2016 ICA Games Division preconference 'Just Games', Tokyo, Japan: All, A., Nuñez Castellar, E.P. & Van Looy, J. (2016) Assessing the effectiveness of digital game-based learning in a commercial enterprise: A case study.

1. Introduction

Corporate managers are constantly looking for more effective and efficient ways to deliver trainings to their employees, which has led to an increasing interest in technology enhanced learning over the past decades (Short, 2014). Technology-delivered instruction does not require separate training facilities, travel costs for employees and employees/trainers being away from the job, resulting in a more cost-efficient training method for large companies (Joo, Lim, & Park, 2011). Moreover, technology-delivered instruction provides the advantage of convenience and self-paced learning. While benefits of technology delivered instruction have been widely recognized, enthusiasm among employees to use e-learning programs, however, is rather low (Pannese, Cassola, & Grassi, 2005). E-learning is often still related to the passive learning of facts and is not able to engage the learners, resulting in high drop-out (Joo et al., 2011; Pannese & Carlesi, 2007).

Digital Game-Based Learning (DGBL), which refers to the usage of the entertaining power of games to serve an educational purpose could provide a solution to this motivation problem (Prensky, 2001). For this reason commercial enterprises are increasingly investing in the development of games to serve training purposes (Donovan & Lead, 2012; Michaud, Alvarez, Alvarez, & Djaouti, 2012). While a growing number of studies can be found assessing the effectiveness of DGBL in a school and health context, literature testing its effectiveness in a corporate context is scarce, especially studies taking into account the motivational outcomes. This study presents an effectiveness assessment of a digital game-based intervention implemented at a large bank to teach new employees the corporation's client-oriented principles.

1.1. *Interactivity, motivation and learning*

DGBL can be motivating in two ways. Firstly, DGBL can be implemented to 'seduce' the learner by gameplay to allocate his/her attention to the learning content (Ritterfeld, Weber, Fernandes, & Vorderer, 2004). Interactivity is one of the main characteristics of game-based learning resulting in higher attention during the activity and consequently, deeper processing of the content (Ritterfeld, Weber, Fernandes, & Vorderer, 2004). Secondly, DGBL can stimulate intrinsic motivation to engage in the training due to the enjoying experience it provides (Garris, Ahlers, & Driskell, 2002). This means, for instance, that learners wish to finish the game training because it is fun or because they wish to achieve in-game goals rather than

because they are obliged to finish the training. Intrinsic motivation is, in turn, related to higher levels of engagement, performance, higher quality of learning and lower levels of dropout (Ryan & Deci, 2000). Interactivity is, again an important feature of digital games that can stimulate intrinsic motivation (Hwa Hsu, Lee, & Wu, 2005)

While indeed these motivational aspects can be very promising and have been widely recognized in the DGBL field, these all imply that everyone wants to play games and that by the simple act of introducing them, success is automatically achieved. However, DGBL participation can be a result of external coercion, influencing enjoyment of the activity and consequently, learning outcomes (Boyle, Connolly, & Hainey, 2011; Mayer et al., 2014). Hence, in this study we assess whether the motivational mechanisms that underlie DGBL hold true in a corporate context where DGBL is part of a compulsory program. Based on the literature we propose the following hypothesis:

> *H1: Employees will find DGBL more motivating and enjoyable compared to a more passive form of technology-enhanced learning.*

The added value of DGBL is, however, not only related to its motivational power, but its learning mechanisms also fit well within modern theories of effective learning proposed by educationalists and psychologists (Boyle et al., 2011). DGBL provides *experiential learning* opportunities (Rooney, 2012) whereby the learner is not a passive actor, but an active one, learning by doing and reflecting. Hence, we propose the following second hypothesis:

> *H2: Employees instructed by DGBL will score better on a knowledge test compared to employees instructed by a more passive form of technology-enhanced learning.*

A second aim of this paper is to test the feasibility of a procedure that has been developed for assessing the effectiveness of DGBL. This procedure pursues to be applied flexibly across contexts. The procedure has already been tested in a school context (All, Meesschaert, Nuñez Castellar & Van Looy, 2015) and a health context (All, Plovie, Nuñez Castellar & Van Looy, 2016). The present paper represents the feasibility test of the procedure in a corporate context. Based on this validation study, a final version of the procedure will be developed.

2. Methodology

2.1. *Stimulus materials*

2.1.1. Game

The game that was tested has been developed for a large bank in order to teach new employees the bank's basic principles of customer-friendliness in order to improve their loyalty to the bank. The game consists of 5 minigames. The first minigame focusses on client-oriented rules that should be applied before clients are received (e.g., clean office); the second on client-oriented principles that should be applied at the reception (e.g., make eye contact with entering customers); the third on client-oriented principles to be applied when dealing with a client (e.g., empathize with the environment of the customer); the fourth on client-oriented goodbye (e.g., accompany the client to the exit) and the last minigame on client-oriented organization during a day at the office (e.g., follow up on the bank's general mailbox). The game is available to all employees via the online learning platform of the bank, accessible only via the intranet of the bank. The minigames can therefore only be played in the workplace.

2.1.2. Instructional video

For the purpose of this study, an instructional video was developed, using the game and game play as a basis. For this purpose, the screen of the game was captured while being played by the researcher. To make it look more like an instructional video and less like a game, in game-actions were accompanied by texts boxes, explaining why a certain decision is taken or why a certain action is being carried out. For instance, in one minigame, one has to attribute priorities to certain in game events; if a person walks in and at the same time the phone is ringing, one should answer the phone before the third ringtone, one should make eye contact with the customer coming in and after the phone call is finished, the employee should ask the customer how he/she could be of service. When playing the game, one has to drag and drop events based on their priority within a certain timespan. In the instructional video, one sees the events being dragged and dropped based on priority, but a small text box is added next to every event that is being dropped: 'when the phone rings, one should answer within the time of three ringtones', 'While answering the phone, make eye contact with the customer' and 'one you have finished your call, ask the customer how you could help him/her'. Hence, the content treated in the game and the instruction video is exactly the same. The only difference between the two instructional materials is interactivity.

2.2. *Design*

A pre-test post-test control group experimental design was implemented whereby one group had to finish the game training and another group the instructional video training. Considering that the game was developed for cost-efficiency reasons, no 'business as usual' was available as the only training available was the game-based one. Hence, we could not compare the game-based group to a group that received a more 'traditional' intervention as suggested by the procedure. Instead, one group that did not receive an intervention served as a control group. Blocked random assignment (i.e., 'matching') was used to assign participants to conditions. Blocks were created based on age, number of months working at the bank and gender. As prescribed by the procedure (All, Nuñez & Van Looy, 2016), the game was played in the context in which it is meant to be played: during working hours at the employee's convenience.

2.3. *Measures*

2.3.1. Cognitive learning outcomes

Two parallel versions (i.e., same types of questions and difficulty level) of a knowledge test were developed based on the content treated in the games in cooperation with the training manager of the bank. We choose for administering parallel versions pre- and post-intervention (same types of questions and same type of difficulty level), to reduce pre-test influences (Crawford, Stewart, & Moore, 1989; Randel, Morris, Wetzel, & Whitehill, 1992). Test development consisted of 3 iterations, on which we will not elaborate on due to space limitations. The final tests used for the study consisted of 16 questions: 3 ranking questions where different events at work need to be ranked according priority; 4 open ended questions and 9 multiple choice questions. The scoring occurred accordingly: correct ranking yielded 3 points and a correct answer on the multiple choice yielded one point. The open questions accounted for 12 points. The maximum score for the test was 30. One half of the participants received version A and the other half received version B at the pre-test and vice versa at the post-test.

2.3.2. Motivational outcomes

The IMMS -Instructional Materials Motivation Survey- (Keller, 1987) was used to assess motivation towards the instructional method. We based ourselves on Huang, Huang & Tschopp (2010) for the game version of the IMMS. The IMMS consists of 36 items, divided in 4 subscales: attention (i.e., gaining and keeping the learner's attention), relevance (i.e., activities must relate to current situation or to them personally), confidence/challenge (i.e., activities

cannot be perceived as too hard or too easy) and satisfaction/success (i.e., learners must attain some type of satisfaction or reward from the learning experience). The items were scored on a 5-point Likert scale. The interest/enjoyment scale developed by Ryan (1982) was also used in the post-test. The scale consists of 7 items that are rated on a 7 point Likert scale.

2.4. *Participants and procedure*

An e-mail was sent by the training manager with a link to the online pre-test on to all people who had started working at the bank between 1 and 12 months before the start of the study (n = 89). After filling out the pre-test, participants received 6 weeks to complete the training (game or instructional video) on the electronic learning platform. It was not necessary to play/watch all five games/video's consecutively, but they could choose to spread them over several days/weeks. The training manager could retrieve weekly reports on who participated in each mini game/video and provided them to the researcher. One week before the six-week intervention period had passed, the researcher sent a reminder to those who did not finish the game yet, asking them to complete the training considering they would receive a post-test a week later. If they still not had finished the training 6 weeks after the pre-test, the researcher contacted the employees by phone. Once the employees had finished the training, the researcher sent them an e-mail with the link to the post-test.

3. Results

In total, 64 employees participated in the study, of which 20 employees trained themselves with the game, 21 with the instructional video and 23 served as a control group. Table 1 shows that no pre-existing differences exist between the groups regarding age, gender, previous work experience at a bank, game experience (games at least a couple of times a year) or scores on the pre-test, showing successful randomization. Moreover, no differences were found between the two versions of the knowledge test on the pre-test, showing that both tests can be considered equal, $F(1,62) = 1.59$, $p = .21$

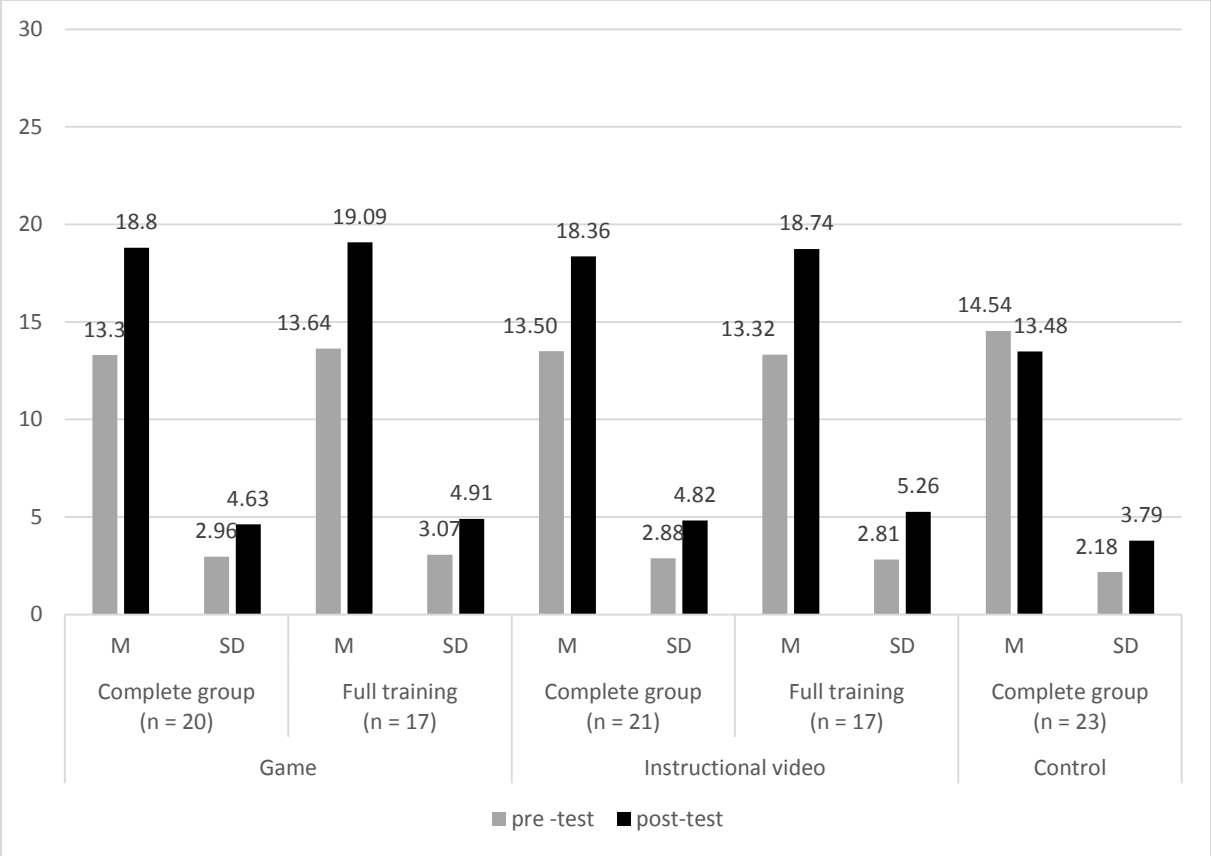
Table 1: Control for balanced groups as a result of randomization (N=64)

	Game (n = 20)	**Instructional video (n = 21)**	**Control (M/SD) (n = 23)**	**F/ Chi²**	**p**
Age (M/SD)	29.95/5.34	29.62/7.40	28.70/5.79	.24	.79
Female gender (n)	13	15	16	.21	.90
Previous professional bank experience (n)	2	4	5	1.11	.57
Gamer (n)	15	13	14	1.85	.40
Pre-test (M)	13.3	13.5	14.54	1.37	.26

Four participants from the instructional video group did not complete all four video's and three participants from the game group did not complete all mini games when filling out the post-test. Hence, we have conducted the analyses twice: once on the complete dataset (n = 64) and once only including the participants that have fully completed the training (n = 57) .

Results show a significant gain from pre- to post-test ($p < .01$) with a large effect size for both the game and instructional video group ($r = .57$ for the complete game group, $r = .51$ for the complete instructional video group, $r = .54$ for those who fully completed the game training and $r = 0.57$ for those who fully completed the instructional video training). The control group shows no significant difference between pre and post-test ($p = .14$). For the complete dataset, the biggest gain from pre- to post-test can be found in the game group ($M = 5.5$, $SD = 4.93$), followed by the instructional video group ($M = 4.86$, $SD = 4.84$). The control group slightly declined ($M = -.1.07$, $SD = 3.38$).

Figure 1: pre- and post-test scores for all groups



An ANOVA on the gain scores shows a main effect of treatment with a large effect size $F(2,61) = 14.90$, $p < .001$, $r = 0.57$. Post hoc Scheffé tests show that the gain of the game and video group is significantly larger than the control group ($p < .001$). No significant differences can be found regarding progress on the knowledge test between the game and instructional video group ($p = .90$).

For the participants that have fully completed the training, we consist of data on when they finished the training. Hence, for those participants, we can conduct an ANCOVA with the time between start and finish of the training and time between completion of the training and post-test as a covariate, allowing us to control for these potential confounding variables. As time between start and finish of the training violates the assumption of independence and the treatment effect, this was omitted from the ANCOVA analysis. More specifically, the instructional video group finished the training in less days ($M = 1.18$, $SD = 3.11$) compared to the game group ($M = 30.65$, $SD = 38.35$), $F(1,32) = 9.97$, $p = .003$. After controlling for time between completion of the training and the post-test, the game group shows an average gain of 5.40 ($SD = 1.19$) and the instructional video group of 5.56 ($SD = 1.19$). Still, no significant

differences can be found between the game and instructional video group, F(1,31) = .001, p = .97. Hence, we need to reject H2.

For the analysis of the IMMS (N = 64), one case was excluded due to non-response on all IMMS items. Here, also no significant differences can be found between the game and instructional video group on the total IMMS score, $F(1,38) = .27$, $p = .61$. When conducting a MANOVA on the subscales, also no differences can be found, F(4,35) = 1.45, p = .24. The scores on the IMMS and its subscales and can be found in table 2. While interpretation is rather difficult considering that the IMMS has not yet been implemented in a study in a corporate context and we have no scores to compare it to, we have found one study stating that instructional material can be considered successful if the average score on the IMMS and its subscales is 3.5 or more (Pittenger & Doering, 2010). If we apply this threshold, the game nor the instructional video can be considered successful. Also, the total score on the IMMS is below the midpoint of 108 for both groups.

Table 2: Mean and standard deviation on Instructional Materials Motivation Survey (N=41)

	Attention		**Relevance**		**Confidence**		**Satisfaction**		**Total score on IMMS**	
	M	SD	M	SD	M	SD	M	SD	M	SD
Game	2.61	0.25	3.22	0.36	2.90	0.18	2.85	0.42	105.76	6.43
Instructional video	2.54	0.32	3.43	0.43	2.84	0.28	3.01	0.53	107.15	10.26
p	0.44		0.12		0.44		0.31		0.61	

For the interest/enjoyment scale also one case was excluded due to non-response on all items. The game group scores on average 3.72 (SD = .52) on enjoyment and the video group scores on average 3.83 (SD = .64) on the 7-point scale. Again, no differences can be found between both instructional groups for enjoyment, F(1,38) = .34, p = .56. Hence, we have to reject H1.

4. Conclusion & discussion

Both the game and the instructional video proved to be effective in terms of learning outcomes, as they increased knowledge compared to only on-the job experience. The interactivity of the game, however, did not add value to learning or motivational outcomes.

Thus, the idea that games are automatically a more motivational alternative for 'passive' technology delivered instruction, does not hold true in a corporate context -in this case. Consequently, the instructional video could in this case be considered as more effective, as the development of an instructional video is typically cheaper. This means that corporations need not always invest in DGBL as similar results can be achieved using (cheaper) more traditional ways of technology-delivered instruction. This is in line with a study comparing the effectiveness of several technology delivered instructions in a military context (Parchman, Ellis, Christinaz, & Vogel, 2000).

A second goal of this study was to test the feasibility of a standardized procedure to assess the effectiveness of DGBL in a corporate context. A main issue we encountered is the impossibility to compare the game with traditional instruction that is currently implemented, as there is none. Not adding a control group to where another educational activity is being implemented in a DGBL effectiveness study, has been criticized by several as non-rigorous research (Clark, 2007; Hays, 2005). While for the present study, we have developed the instructional video, to answer a research question relevant for the e-learning field, this was not at the request of the company. For the company, there is no added value in developing a 'control instruction condition' just for the sake of research. Hence, we would like to refute the necessity of a control group where another educational activity is implemented if there is no other current method to compare it to. We would however, suggest, to make meaningful comparisons. In the present study, the question the training manager had was simply 'does the game help new staff gain insight in client friendly principles?' In this case, comparing to a group that does not receive extra instruction is not meaningless, as it looks at the added value the game provides compared to on the job experience. In this case, 'business as usual' could thus simply be no extra instruction.

A second issue we have encountered with the procedure, is the suggestion to implement the game in a context for which it has been developed to improve external validity. This quasi-experimental design was, however, far from ideal. While almost everyone filled out the pre-test, a major issue in the present study was motivating the employees to start playing the game, even though it was compulsory. The researcher had to track activity of every individual participant, following up on whether or not they had already started playing the minigames/watching the instructional videos. Subsequent e-mailing and calling participants several times to finish the training reduced external validity as this is not common practice in the corporation.

Related to this, if we would take the cost of *monitoring* whether or not the employees followed the training into account and following up on those who did not, we can put the efficiency rationale behind technology delivered instruction and game-based learning -in this case- in doubt. The lack of motivation to start the training is also detected on a broader scale within the company, as only 200 of 8000 employees have already played the game. This lack of motivation to start playing is unlikely to be related to individual underlying reasons such as technology skills, game skills or attitudes towards games considering it was as difficult to motivate the participants in the instructional video group, which did not require any of these skills. Hence, a more plausible explanation might be related to the format of the training. Making the training only accessible at the office and consequently, during working hours may have impeded employees to play the game. Time management has indeed previously proven to be an issue for employees to actually use e-learning programs (Joo et al., 2011). Other impeding factors are the lack of social interaction on the platform (Short, 2014), the lack of supervisory support and, related to this, lack of incentive to engage in e-learning programs (Joo et al., 2011). The non-engagement to start the training might thus be related to the lack of a meaningful learning context (De Freitas, 2006). While DGBL was not successful in solving engagement issues that are encountered in more passive e-learning approaches, it has the potential to tackle these issues in a way passive e-learning initiatives cannot, that is, by using game mechanics. For instance, a simple score board in the game, creating competition between colleagues could provide a solution for the lack of social interaction and incentive. Hence, the reason why the game did not add value to the instructional video, is that the motivational game features may not have been used to their full potential.

To conclude, the success of games as instructional medium in a distance self-paced learning context is not only related to the question ‘If learners play it, does it improve motivation, learning outcomes and/or cost-efficiency?’ but also ‘does it succeed in getting learners to actually start playing?’. Further research should thus not only focus on whether DGBL is effective and which in-game elements make DGBL effective, but also on which implementation methods or context variables motivate employees to actually start the game-based training.

5. Limitations

Due to practical limitations –we could only include new employees that had started working at the bank- we had a small sample size. A second limitation is that intervention period possibly

confounds our results, as it was significantly different between the instructional video and game group.

6. Statement on potential conflicts of interest, open data and ethics

This study was set up as part of a Ph.D. project, aiming to validate a standardized procedure for assessing the effectiveness of Digital Game-Based Learning in several contexts. Consequently, this work has been independently conducted and analyzed by the researchers. The authors state that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

The employees were informed that they were taking part in a study assessing the effectiveness of several learning media that was part of a Ph.D. project. We also made clear in the e-mail that was sent out and again in the pre-test that their personal information such as their name and surname was only available for the researchers in order to link the pre-test to the post-test and that once all data was gathered, their names would be deleted and replaced by a participant number. At the end of the e-mail we again stressed that their data would be treated with confidentiality and would be processed in an anonymous manner.

The anonymized dataset has been added as a supplementary file to this manuscript and is thus publically available.

PART 3.

EPILOGUE

CHAPTER 8.
General Conclusion and Discussion .

The aim of this Ph.D. project has been to develop a standardized procedure for assessing the effectiveness of digital game-based learning (with a primary focus on games that target cognitive learning outcomes) by firstly determining study design characteristics of published (quasi-) experimental studies on the effectiveness of digital game-based learning (DGBL) aimed towards cognitive learning outcomes (chapter 2). Secondly, we conceptualized and operationalized effectiveness of DGBL (chapter 3). Thirdly, we defined best practices for assessing the effectiveness of digital game-based learning (chapter 4) in order to finalize a first version of the procedure. In a second stage, we tested the feasibility of the procedure by means of experimental studies using this procedure as a guideline in order to further optimize it (chapter 5-7).

1. Summary of main findings

1.1. *Issues in published effectiveness studies of DGBL*

A systematic literature review was conducted of published DGBL effectiveness studies of games that primarily target cognitive learning outcomes (chapter 2). Results pointed out three issues. Firstly, heterogeneity exists among separate studies. While indeed one could say that this heterogeneity is a logical consequence of the variety of different game genres and topics integrated in DGBL, the main causes of this heterogeneity significantly impact results. More specifically, the three main issues causing heterogeneity are different activities that are implemented in the control group, different measures that are used to assess effectiveness and different statistical techniques for quantifying learning outcomes. For instance, comparing the experimental condition to a no-activity control group or a control group where exactly the same content is treated using another medium, will yield very different results (Stewart et al., 2013). Also problematic is that motivation is considered a key concept in DGBL (Garris, Ahlers, & Driskell, 2002), but that it is only seldom assessed. Moreover -even when assessed- different types of motivation are measured across studies (e.g., motivation towards the instructional materials, enjoyment and general motivation for school/a course), demonstrating the need for a clear delineation of what one desires to achieve regarding motivational outcomes of DGBL. Lastly, different measures used for analysis and different statistical techniques for quantifying outcomes can again significantly impact results: only comparing post-test scores or comparing progress, for instance, will yield different results.

A second important issue with certain studies is that they use a suboptimal study design, leaving the door open for confounds, variables which do not relate to the manipulation but which have in previous studies shown to significantly impact learning outcomes. More specifically, the addition of extra elements to the intervention (i.e., required reading, debriefing session, etc.) (Sitzmann, 2011; Wouters, Oostendorp, Boonekamp, & Spek, 2011) and the presence, type (i.e., familiar vs unfamiliar) and role (i.e., supervision, procedural help or guidance) of the instructor during the intervention (Brom, Šisler, Buchtová, Klement, & Levčík, 2012). Also, often the same tests are used pre- and post-intervention, which can lead to practice effects (Randel, Morris, Wetzel, & Whitehill, 1992). These confounding elements make it difficult -if not, impossible- to know whether the same beneficial effects would have been found without these elements.

A third and final issue with DGBL effectiveness research is reproducibility. Often very little information is given about how the interventions were implemented (e.g., who was present? In which context was the game played? Was gameplay individual? Etc.), how sampling occurred, how similarity was attained between experimental and control group, which tests were implemented and if tests had been developed by researchers themselves, how these were developed. Detailed information on these elements is indispensable, however, in order to gain insight in whether the gains that are reported are a consequence of the different methods or due to other circumstantial factors that differed between conditions (Randel, Morris, Wetzel, & Whitehill, 1992).

In sum, the systematic literature review has pointed towards the need for guidelines or at least the formulation of some minimal requirements for conducting DGBL effectiveness studies in order to make study designs more comparable and facilitate the comparison of results across studies. This would allow researchers to make more general claims on effectiveness of DGBL and potentially strengthen the field as a whole.

1.2. *Conceptualizing and operationalizing DGBL effectiveness*

In chapter 3 we have developed a conceptual framework for effectiveness of DGBL. For this purpose, we have conducted a requirements analysis by means of 3 focus groups with 3 relevant stakeholder groups. We have rooted our conceptualization in Social Cognitive Theory (Bandura, 1986). More specifically, we have used Bandura's concept of agency as a theoretical framework for defining DGBL effectiveness. Agency refers to humans' ability to influence their own behavior through intentionality, forethought and self-regulation by self-reflectiveness

and self-reactiveness about their behavior (Bandura 2001). This means that individuals are capable of evaluating their own behavior (i.e., self-evaluation, self-reflectiveness) through observation of that behavior and the associated outcomes (i.e., self-observation). Based on this evaluation, behavior is (dis)continued or altered (i.e., self-reactiveness). This evaluation of behavior occurs based on goal setting which refers to objectives one wished to attain by performing a certain behavior. Hence, outcomes one desired to attain through a particular behavior serve as a benchmark against which to judge effectiveness (Bandura, 2001). If we apply this to digital game-based learning, the evaluation of its effectiveness will be against outcomes one desires to attain by implementing DGBL. Hence desired outcomes of DGBL are considered as the cognitive component influencing behavior, which is implementation of DGBL. Thus desired outcomes of the implementation of DGBL serve as a benchmark for evaluating its effectiveness.

Our results indicated that effectiveness of DGBL is a multidimensional construct consisting of three categories of desired outcomes: learning, motivational and efficiency outcomes. For each category, several indicators can be used. For learning outcomes this can be an increased interest in the subject matter, performance (i.e., on a knowledge test) and transfer, referring to applying required knowledge in the real world. Motivational outcomes can refer to simply creating a more enjoyable learning experience compared to current instructional methods or to make learners more motivated to learn using DGBL. However, motivation is not a stand-alone reason to implement DGBL, it should still be related to similar learning outcomes compared to current instructional media. Efficiency outcomes refer to reducing time for teaching a certain content matter or providing a more cost-effective solution for teaching a content matter to a certain group of learners. Similar to motivation, higher efficiency outcomes compared to traditional instructional media are not a stand-alone reason to implement DGBL. Higher efficiency outcomes should still be related to similar learning outcomes achieved by more traditional media.

By developing a conceptualization of effectiveness based on desired outcomes of DGBL, it indirectly relates it to use. More specifically, if DGBL succeeds in generating outcomes considered relevant for the stakeholders and these outcomes are empirically validated, this will support adoption and usage of DGBL.

1.3. *Best practices for assessing DGBL effectiveness*

In chapter 3 we have defined best practices for assessing the effectiveness of DGBL, based on semi-structured interviews with experts on intervention research coming from the field of psychology and educational science. In this chapter, we have detected several potential areas for improvement in the field of DGBL effectiveness research: the implementation of the intervention and the methods employed to assess effectiveness. Regarding implementation of both the interventions in the experimental and control group, several practices were defined that are preferably avoided during the intervention in order to reduce confounds (such as guidance by the instructor, extra elements that consist of substantive information) and which elements could be allowed (e.g., procedural help, training session). Moreover, variables on which similarity between experimental and control condition should be attained were determined (e.g., time exposed to intervention, instructor, day of the week). With regard to the methods dimension, proposed improvements related to assignment of participants to conditions (e.g., variables to take into account when using blocked randomized design), general design (e.g., necessity of a pre-test and control group), test development (e.g., develop and pilot parallel tests) and testing moments (e.g., follow up after minimum 2 weeks). In sum, this chapter provides best practices that cover all aspects of the study design. While several suggestions have previously been made regarding research design of DGBL effectiveness studies (Brom et al., 2012; Serrano-Laguna et al., 2013), these do not cover all aspects of the research design, such as aspects for which similarity between subjects should be attained between experimental and control group, instructor role and implementation of the intervention.

1.4. *Empirical findings on DGBL effectiveness*

Chapters 5-7 deal with experimental studies aiming at the same time to further improve and validate the proposed best practices and to make an assessment regarding DGBL effectiveness in a number of different contexts. In our first feasibility study in a school context (chapter 5), an interesting finding was that while at the first post-test no significant difference could be found between the group that had learned English vocabulary using DGBL and the group that had received a traditional class by the teacher; at the second post-test -three weeks later- the group that had received the traditional class outperformed the group that was instructed by DGBL. Thus, in the longer term the traditional class proved to be more effective.

This supports previously made claims on short term effects in computer-based learning (Clark, 1983). Moreover, a debriefing session did not add value regarding learning and motivational outcomes to the game-only condition. This goes against part of the literature as it has been suggested that a debriefing is indispensable in digital game-based learning (Crookall, 2014; Pivec, 2007; Van Der Meij, Leemkuil, & Li, 2013; Kriz, 2008). This has raised a number of questions regarding delineation of DGBL characteristics that require a debriefing. More specifically, nuances regarding game type, complexity of learning content, explicit/implicit learning goals, game characteristics and possibly other factors need to be explored.

In our second feasibility study in a health context (chapter 6) we found that adding a pre-test to an effectiveness study of DGBL can influence results as pre-test sensitization only occurs in the group that received a slide-based lecture. More specifically, the participants that received a pre-test before the slide-based lecture had significantly higher post-test scores than the slide-based group that did not receive a pre-test before the lecture. In the game group, no significant differences could be found between those participants that received a pre-test before the game-based training and those that did not. This makes comparison of DGBL and more traditional classes in a pre-test post-test control group design rather difficult, as post-test scores of the traditional class might be positively biased. However, the fact that pre-test sensitization does not occur in the DGBL group also confirms the effectiveness of DGBL, as the interactivity of the game required them to be attentive, regardless of whether they received a pre-test or not before the DGBL intervention, confirming the motivation paradigm proposed by Ritterfeld and Weber (2006). Furthermore, both game groups still outperformed the slide-based group that received a pre-test before the lecture, conforming the higher effectiveness of the game. This study has thus also shown the added value of conducting a Solomon 4-group design in the context of DGBL.

In our third feasibility study, which took place in a corporate context (chapter 7), the interactivity of the game was found not to add value to a passive instructional video that delivers exactly the same content. The motivation rationale behind DGBL did not hold true in this case pointing to the need for careful consideration as to where to use interactive content.

In sum, in two of the three feasability studies, DGBL did not prove to more effective than more traditional, ‘less engaging’ instructional media. Does this mean that this supports critics’ statement that positive findings in favor of DGBL might be attributed to less rigorous design elements? (Clark, 2007; Hays, 2005). This is something that needs to be further investigated in the future and is increasingly being integrated in meta-analyses. For instance, in a recent meta-analysis by Clark and colleagues (2015), impact of study quality was investigated.

Results have shown here that the effect size decreased when taking the quality of the research design into account. However, quality of research design was defined based on the quality of the control condition and whether or not a randomized controlled design was used. A more nuanced view on quality of research designs (e.g., confounds during the intervention, similarity of conditions, instruments used, integration of a follow-up study, etc.) should thus be taken into account in the future. Based on our findings, it is possible that some studies jump to conclusions or do not provide information in order to get more nuanced view as a reader of possible alternative explanations for their outcomes. This has also been broached by other authors (Clark, Tanner-Smith, & Killingsworth, 2015; Sitzmann, 2011).

Another question related to the 'disappointing' outcomes of our effectiveness studies is linked with the 'can media influencing learning? discussion referring the the Kozma vs. Clark media debate (Clark, 1994; Kozma, 1994) we have briefly touched upon in the introduction. In the context of DGBL we do believe game-characteristics can influence learning outcomes. For instance, the interactivity of the fire safety training game in our second feasibility study increased attention and consequently, learning outcomes. Also, in our first feasibility study, both gaming groups (with and without debriefing) found the game more fun; which is also a desired outcome of DGBL related to the game as a medium. However, we do believe that often, game mechanics are not used to their fullest potential and DGBL does not always succeed in finding the right balance between entertainment and instruction. This is also confirmed by equal motivation for learning through the instructional media in our second (chapter 6) and third feasibility study (chapter 7).

2. Reflection

2.1. *What is effectiveness of Digital Game-Based Learning?*

What became clear in this dissertation is that effectiveness of DGBL is a complex construct. The several dimensions and sub dimensions defined in chapter 3 can be approached in different ways. An important distinction that comes forward in this dissertation is the difference between absolute effectiveness and relative effectiveness. What type of effectiveness will be required, will ultimately depend on the research question.

Absolute effectiveness refers to the simple question: does DGBL succeed in achieving its predefined goals? This thus primarily refers to learning outcomes and refers to the investigation of progress regarding those learning outcomes -which could be increased interest, performance and/or transfer- as a result of the game. Hence, this requires an analysis from pre- to post-test.

It is still recommended to also have a control group, to investigate whether differences between pre- and post are a result of the mere lapse of time (Campbell, Stanley, & Gage, 1963). Interpretation of motivational outcomes is more difficult as this is a post-intervention measure. Here, only descriptive analysis of the scores is possible.

Relative effectiveness refers to the question: is DGBL similar or even better compared to the other instructional media? Here, preferably, the media that are currently implemented to teach a certain subject matter are used. With relative effectiveness, all dimensions of DGBL effectiveness defined in chapter 3 are considered relevant. Note that when using the relative effectiveness approach all parameters concern a judgment of relative worth, comparing the outcomes to the current instructional medium used for teaching a particular content matter, implying the need for a control group where another educational activity is implemented.

The distinction between absolute and relative effectiveness is an important one to be made, as DGBL effectiveness studies that do not implement an educational activity in the control group, are often criticized as not being rigorous (Clark, 2007; Clark et al., 2015; Hays, 2005). In some cases, however, there is no control group available to which the content matter is taught in a more traditional way. For instance, a training manager can choose to develop a new training immediately using DGBL for cost-efficiency reasons, based on a cost-benefit analysis. This was the case in our third feasibility study (chapter 7). Here, the training manager simply desired to know whether the game they had invested in actually succeeded in achieving its goal to guide further investment decisions regarding game-based trainings. In this case, there is no use in creating a more traditional training just for research purposes. A control group where no educational activity was implemented was useful here and was not a 'strawman' comparison which DGBL studies with a no educational activity control group are often criticized for (Clark et al., 2015). More specifically, the results of this study provided us and the training manager with insight in whether the game provided an added value compared to only on the job experience of working with customers. .

2.2. *Trade-off between ecological validity and control*

While in experimental research, control of as many elements as possible that might influence results is an important aspect, it can also be problematic in the context of DGBL. A first reason for this is the complex environments in which DGBL is often being implemented, such as natural collectives in which one does not always have control over observed and unobserved variables. In our first feasibility study, for instance, we had assigned participants to

conditions by randomizing on the subject level. However, this implied that pupils from different classes were divided into new groups (2 classes from the fourth and 2 classes from the fifth grade were divided over 3 conditions), potentially threatening ecological validity. This, however, resulted in successful randomization, increasing internal validity as there were no pre-existing differences on the pre-test scores of the English vocabulary test. In our second feasibility study, we randomized on a group level. However, in this study randomization failed, resulting in pre-existing differences in the pre-test scores of the fire safety knowledge test, threatening internal validity. Moreover, it resulted in some issues for analyzing our Solomon 4-group design. Hence, in our third feasibility study, we have used blocked random assignment to assure similarity between conditions. Blocks were created based on age, number of months working at the bank and gender. This has resulted in a successful randomization and thus no pre-existing differences on the pre-test scores regarding client-oriented principles of the bank.

Another reason why control is not always possible or desirable is that the main rationale behind implementing games as instructional tools is one of motivation (Garris et al., 2002). Hence, implementing a game in a controlled lab setting would provide us with limited insight in motivational outcomes as this is a highly artificial environment. Giving in on control for ecological validity can in turn get in the way of ecological validity. For instance, in our third feasibility study (chapter 7), the game we tested was meant to be played by the bank employees at their own convenience, during office hours. Hence, the decision to provide the employees with 6 weeks to finish the training at their own convenience, as this is the way it would occur in real life. This way, motivation for the instructional material was more ecologically valid, as they would play when they would feel like it or when they had some time available. However, we had to contact the employees several times to start or finish the training, that this in turn affected ecological validity; since in real life there would be no one nudging the employees. In hindsight, we still believe this was the best way to conduct this study as it gave is some valuable insights in providing meaningful contexts for DGBL that resulted in some recommendations for the bank to further optimize their DGBL trainings in the future.

The complexity of DGBL effectiveness also results in a trade-off between control and ecological validity. One of the main critiques regarding research rigor in DGBL effectiveness studies is differences in instructional time between the experimental (game) and control (traditional education) groups (Clark, 2007; Randel, Morris, Wetzel, & Whitehill, 1992). However, in our second feasibility study (Chapter 6) it has become clear that this is not always desirable. In a context where learners are paid employees, keeping instructional time equal for the experimental and control group is often difficult, as a reduction of training time and as a

result, higher cost-efficiency is often a desired outcome in these contexts. Hence, keeping instructional time equal is incompatible with the efficiency outcomes of DGBL. In such cases, instructional time should be treated as an outcome and research should focus on investigating whether learners learn as much or more in less time using the game-based method.

To summarize, a balance between ecological validity and control is thus best achieved by firstly increasing internal validity by reducing the influence of confounding variables during implementation of the intervention(s) as much as possible. For instance, in our first (chapter 5) and third feasibility study (chapter 7), we have developed parallel versions of (i.e., same types of questions and same difficulty level) of a knowledge test that we implemented pre- and post-intervention, in order to reduce the confounding practice and pre-test sensitization effects of implementing the same test pre- and post-intervention. As became clear in our second feasibility study (chapter 6), a pretest can bias learning outcomes. More specifically, we have shown that pre-test sensitization took place in the control group that received the fire safety training by means of a slide-based lecture. In the game group, no such effect was found. As mentioned in chapter 6, this is a relevant methodological issue in the field of DGBL as the learning outcomes in control groups receiving more traditional lectures might be positively biased, while in DGBL these probably represent more 'true' scores, as interactivity of the game requires them to process the content in order to finish the training. Hence, it becomes tricky to compare these groups in a pre-test post-test design; especially considering that often non-significant differences are found between DGBL and traditional classes (Clark et al., 2015).

Internal validity can also be increased by keeping potential sources of variability other than the experimental variance (e.g., date and time of the intervention, amount of support provided, content, etc.) equal between experimental and control group. Hence, it is important to report on how this similarity was attained for a) the interventions (what efforts were made to keep interventions equal? Which elements differed between the interventions of experimental and control groups?) and b) relevant participant characteristics. For instance, in all our feasibility studies we made sure that exactly the same content was treated in experimental and control conditions and aimed towards the presence of the same instructor during both interventions. Also, we have aimed to keep experimental and control group as similar as possible regarding relevant observable participant characteristics, such as game experience, age and gender.

If there is an imbalance between groups regarding relevant participant variables that might influence the outcomes, this could be added to the analysis in order to take this difference into account. For instance, in our second feasibility study, we have added pre-test scores as a covariate in order to control for initial differences on the pre-test scores. An imbalance

regarding intervention characteristics could also be added in the analysis. For instance, in our third feasibility study, we had added time between completion of the training and post-test as a covariate, in order to control for this potentially confounding variable.

External validity can be maximized by ensuring similarity between elements present in the real world implementation environment and the implementation for the effectiveness assessment, such as implementation in a context in which the game is intended to be used, implementation in natural collectives such as existing class groups (i.e., randomization on a classroom level or blocked random assignment), the presence of a familiar teacher in a classroom, the provision of procedural help, etcetera..

2.3. *Practical context of DGBL effectiveness studies*

While our guidelines developed in chapter 4 are based on academic expertise, some issues may arise when implementing them in real life. In our first feasibility study we have shown that the addition of a follow-up test is necessary to get a better indication of learning gain, even if it is already after three weeks. Longer term follow-up studies can certainly be interesting, especially for interventions that have been implemented for a longer period of time. However, in practice, it is often difficult to gather this type of data. For instance, in our second feasibility study (Chapter 6), we implemented the follow-up study after one year. This was difficult for two reasons. Firstly, not everyone that had participated in the initial study still worked at the hospital. Secondly, since the end of the initial study a year before, the game was available on the hospital's intranet, for employees to play in order to revise the fire safety training. Consequently, the following things happened: a) people that were in the control condition and did not get to play the game, had now played the game between the end of the study and the follow-up; b) people from the game-based group replayed the game during the end of the study and the follow-up and c) people from both groups decided to revise the safety training using the game before they came to fill out the test at follow-up, even if we had stressed not to revise the course in order for us to investigate longer term effects. Consequently, we did not have enough participants left that were eligible for the data-analysis of the follow-up study.

Another example is that, based on our second feasibility study (chapter 6), it has become clear that a Solomon 4-group is a desirable design to conduct when assessing the effectiveness of DGBL, in order to have a more nuanced view on the results and make more strongly supported claims on DGBL effectiveness. Hence, one of our recommendations in our procedure could be to always conduct a Solomon 4-group design. This, however, requires a

doubling of the sample size and might thus be practically problematic in certain situations. For instance, a game developer might work together with a researcher to test the effectiveness of its product, but might have limited budgets and time for this type of study. So, in our procedure we briefly state that this is the most desirable design to get insight in the effects of the pre-test, but that an alternative solution is to implement parallel test versions pre- and post-intervention.

3. General conclusion

The aim of this Ph.D. was to develop a standardized procedure for assessing the effectiveness of DGBL which primarily aim towards cognitive learning outcomes. Based on the studies included in this dissertation we demonstrated that a more systematic and streamlined approach is possible. Firstly, a more streamlined approach on outcome measures is now possible based on our conceptualization and operationalization of DGBL effectiveness brought forward in Chapter 3. Secondly, a more standardized approach regarding actual study design characteristics is possible up to a certain point. For instance, when dealing with DGBL aimed towards cognitive learning outcomes, a minimal requirement is the addition of a pre-test and a control group and the addition of a follow-up test -even if it is already after two weeks. A context of play that is representative for a real world implementation of play is also a relevant aspect of DGBL effectiveness studies due to the fact that motivation is also considered an outcome. Reducing confounds during the intervention should be done by at least assuring there are no extra elements added to the game-based intervention, reducing help during the game-based intervention to procedural help and aim to use parallel tests pre- and post-intervention when using tests developed by the researchers.

However, for some study design aspects, complete standardization is not possible and will ultimately depend on the research question and needs of the people for whom the study is conducted. For instance, whether or not an educational activity should be included in the control group will depend on what is already available on the content matter treated in the game and consequently, whether absolute or relative effectiveness is assessed. Also, the trade-off between experimental control and ecological validity will ultimately be made by the researcher. Important here is the accurate reporting by researchers in order to provide readers with a more nuanced view on factors potentially influencing the outcomes of interest.

Finally, this dissertation has shown that a more standardized approach for DGBL effectiveness assessment is not only possible, but also required. For instance, the concept of effect sizes has become increasingly important in social science research. However, it is not a common practice in DGBL effectiveness studies to report effect sizes. Hence, there are no assumptions on the magnitude of the effects of DGBL. This is however important for future directions of the field to firstly be able to conduct power analyses beforehand to determine the sample size required for conducting effectiveness studies. Secondly, effect sizes can be relevant in order to make claims on a more general level with regard to DGBL effectiveness. However, effect sizes will depend on the research design implemented (i.e., outcome measures used, activity in control group, whether or not a pre-test was implemented, etc.). Lastly, to make claims on DGBL effectiveness and in order to investigate which features of DGBL contribute to its effectiveness, meta-analyses are increasingly being conducted. Studies to be included in these meta-analyses, should, however, have a similar research design (Higgins, 2008), again showing the need for a more streamlined approach.

4. Implications

4.1. *Scientific community*

This dissertation has provided a methodological contribution to the field of media-effects research by defining, implementing and reflecting on best practices for the assessment of the effectiveness of digital game-based learning. During four years we have been able to thoroughly study, test and document a variety of different possibilities regarding every aspects of the study design for assessing the effectiveness of DGBL primarily aimed towards cognitive learning outcomes and advantages and disadvantages related to those options. Most researchers do not have the time to explore all possibilities when having to conduct these studies. Moreover, researchers conducting effectiveness studies on DGBL come from a variety of backgrounds with different research traditions, and do not always have a background in the literature on experimental research. Hence, the added value of this Ph.D. is the reflection on study design characteristics in chapter 2 and best practices presented in chapter 4, which can serve as an information source for researchers working out the study design of an effectiveness assessment of DGBL. Moreover the extensive reflections of several study design characteristics in our feasibility studies can further provide them with information for working out their study design.

An added value of this dissertation for the field of DGBL is the conceptual framework provided in chapter 3. Currently, there was no clear conceptualization and operationalization of

DGBL effectiveness, resulting in different outcome measures that have been used as indicators for effectiveness. Our framework has filled in a need in the DGBL field for a conceptual model that provides a general evaluation framework for assessment of DGBL which can be applied flexibly across contexts.

Another added value of this dissertation. for the field of DGBL is that we tested the feasibility of the procedure by implementing it in a number of real-world effectiveness studies. Hence, the last version of this procedure is not solely based on academic expertise but also takes into account constraints that can occur in real life and has taken into account possible solutions, which have been discussed in the case studies and this final chapter.

Lastly, the study design issues brought forward in chapter 2, the conceptual framework in chapter 3 and the best practices in chapter 4 could also serve as an information source for the development of experimental procedures for other computer-based instructions with a primary focus on cognitive learning outcomes, such as more traditional e-learning programs or educational interventions using virtual and augmented reality.

4.2.*Game developers*

Our conceptual framework in chapter 3 is also a valuable tool for game developers as it provides desired outcomes of relevant stakeholder groups and thus elements to take into account when developing DGBL. The framework can thus also be used as a communication tool between them and clients who order the development of a game. In this way, game development can be fine-tuned to the needs of the client. By using this framework for development, implementation of DGBL is further stimulated as it is rooted in social cognitive theory according to which the more outcomes are desired for the client(s), the higher the likelihood of implementing DGBL.

A second added value for the game industry is that they could have their DGBL interventions be tested using a procedure which has been developed and fine-tuned over a course of four years. When their game proves to be effective, they can refer to a procedure which has independently been developed in the academic community, giving them more credibility. Demonstrating effectiveness is an important factor in the decision making process for adopting or implementing DGBL (Bardon & Josserand, 2009). Hence, this could positively influence uptake of their games. This can in turn have as a result that more game developers let their educational games be tested, resulting in higher quality DGBL products and in turn, a higher quality DGBL industry.

A last, more long-term added value of our procedure is that by streamlining the effectiveness studies more, the possibility is created for conducting meta-analyses where game-characteristics could be taken into account and linked to outcomes, such as conducted by Desmet and colleagues (2014) for health promotion games. This way, researchers can investigate which game characteristics (e.g., competition, narrative, etc.) influence effectiveness and can provide recommendations to the game industry for future DGBL game development. This again can result in more high quality, effective DGBL products.

4.3. *Societal Impact*

Considering that more public resources are allocated to DGBL (e.g., Game Fund in Belgium, H2020 calls), valorization of effectiveness of products that are the result of these funding opportunities is important in order to help decide whether the investment made was a good policy choice. Again, conducting effectiveness studies using the procedure as a guideline, can provide a more neutral way of conducting these studies and consequently, more credibility. Potential adopters can also benefit from the procedure, as it can give them insight in which products have been tested and whether they are effective. Especially given the fact that consumers vouch for the largest part of the global revenues of DGBL in 2014 (Adkins, 2015). This would allow them to make a more informed choice when buying DGBL products, similar to PEGI labels providing them with more information regarding 'negative' characteristics of the game, such as violence and sexual content.

5. Limitations and further research

5.1. *Conceptual limitations*

A first limitation relates to the conceptualization of digital games. In the introductory chapter and throughout this thesis, we have conceptualized digital games as …

> A rule-based formal system with a variable and quantifiable outcome, where different outcomes are assigned different values, the player exerts effort in order to influence the outcome, the player feels attached to the outcome, and the consequences of the activity are optional and negotiable (Juul, 2003, p.5).

However, in a lot of cases, this can also apply for simulations that do not really have other gaming elements than adding scores to in-simulation actions. Pannese and Carlesi (2007) have indeed pointed towards the fact that DGBL often comes down to e-simulations, where 'game'

usually mainly means interactivity. Similarly, the game we have tested in Chapter 6 can be more considered as an e-simulation with the addition of a scoring element than they can be considered actual games. Also the game in chapter 5 can be considered more as an interactive story with certain gaming elements than it can be considered an actual game. This, however, relates to a 'gap' on a larger scale. In our search for games to be tested in the feasibility studies, we did not often come across educational games that truly could be considered as games. Hence, there is definitely room for improvement in the DGBL industry in order to use the motivational power of digital games to their full potential in order to further improve effectiveness.

A second conceptual limitation is related to our conceptualization of DGBL effectiveness (Chapter 3). More specifically, in our user requirements analysis we have defined three stakeholder groups: the operational working area (i.e., game developers and researchers), the wider environment (i.e., stakeholders on a governmental level) and the contacting business (i.e., potential adopters of DGBL). For the latter category, we have decided to focus on intermediaries such as teachers, principals and HR managers. The rationale behind this was that the adoption or implementation decision to implement DGBL is typically made on a higher level (Boyle, Connolly, & Hainey, 2011). However, a recent research report by Ambient insight (Adkins, 2015) has revealed that consumers themselves account for the largest part of the global revenues in DGBL. Hence, extension of the conceptual framework based on desired outcomes of for instance parents who buy educational games for their children and individuals who decide to learn a certain content matter using DGBL is required.

5.2.*Methodological limitations*

A first methodological limitation of this dissertation is that we have solely focused on a procedure that uses an experimental approach. While the aim of this Ph.D. was to develop a procedure for assessing effectiveness of DGBL by means of experimental research, and thus conduct quantitative summative evaluations; the need for extra qualitative data became clear in all three feasibility studies. In the first effectiveness study (Chapter 5) interviews or observations might have given us more insight in to why the debriefing did not add value to the game. In the second effectiveness study (Chapter 6), structured observations could have provided us with more insight on attentiveness during the interventions. This way, we could test our post-hoc explanation stating that pre-test sensitization does not take place in the game group as a result of the interactivity of the game, requiring learners process the content, regardless of whether they received a pre-test or not. In our last feasibility study (chapter 7),

interviews with the participants could have provided us with more insight in why the interactivity of the game did not add value to the instructional video and why it took the participants so long to actually start and finish the training. Answers regarding these questions would have allowed us to make more concrete recommendations to the institutions regarding the format of their training. In sum, applying the procedure to conduct experiments allows us to know if a game is effective or not. The addition of qualitative research would allow us to get an indication as to why it was effective (or not) in order to provide recommendations to further optimize the product or provide recommendations regarding implementation of the intervention.

A second methodological limitation of this Ph.D. is that we were not able to test the feasibility of all aspects of the procedure. For instance, testing a game that is aimed towards home usage and related to this, test the influence of the tech savviness of the parents on the outcomes or adding certain random effects such as teacher and classroom influences. Similarly, we were not able to validated all learning outcomes presented in Chapter 3, such as situational interest and transfer.

Lastly, we did not take enough into account mediating variables that can influence media-effects in our feasibility studies. We did keep relevant characteristics equal in experimental and control groups (e.g., game experience, age, prior knowledge, years of previous experience at the bank, etc.) to reduce the influence of these potential confounding variables. Nonetheless, this way, we can only state whether the tested games proved to be effective or not. It would also be of significant value to identify intervening variables, referring to underlying mechanisms that enhance effectiveness. This can be related to both individual characteristics of the learners themselves (e.g., SES, age, ability) and the interaction between the individual and the game (e.g., involvement in the narrative, identification with the avatar). This way, we can define for whom and under what circumstances DGBL proves to be effective. However, in order to make such claims on a larger scale –by means of meta-analyses-, a more standardized approach for conducting effectiveness studies is required.

5.3. *Further research*

Besides the suggestions for further research presented above, we have three more general interesting venues for further research in the field of DGBL. A first one is that while we have set minimal requirements for standardization of effectiveness assessment of DGBL aimed towards cognitive learning outcomes, further fine-tuning is recommended. For instance, the

development of validated questionnaires for assessing certain outcomes, such as enjoyment and transfer, specifically for the field of DGBL. Moreover, several aspects need to be further investigated, such as whether the simple act of adding a pre-test or the content of the pre-test results in pre-test sensitization.

A second suggestion for further research is to extend the idea of a standardized approach for conducting effectiveness research beyond digital games that primarily aim towards cognitive learning outcomes to skill-based and affective (attitudinal and behavioral change) learning outcomes. A third interesting venue for further research is, testing different implementation methods which has a significant practical value as this can yield recommendations regarding implementation to further optimize DGBL effectiveness A third interesting venue is the development of a 'light' version of the procedure. The procedure in its current format is aimed more towards researchers. A light version could for instance be interesting for game developers that want to do in house effectiveness tests of preliminary versions of games in order to get a grasp on their effectiveness and optimize the design. This can further bridge a gap between research and development.

6. Final considerations

During the course of my Ph.D. trajectory I had the chance to meet a lot of different stakeholders related to the field of DGBL: researchers, developers, teachers and principals, training managers, parents, and people that are involved in DGBL on a policy level.

It has become clear to me that the academic research field of digital game-based learning and the game industry are often two separate entities. Typically, effectiveness of DGBL is assessed either -once the game has been developed- because it is funded by public resources and valorization is obliged or they are conducted voluntarily by DGBL researchers for their own research. Game developers engaged in DGBL typically develop games based on gut feeling, often working with budgets that do not allow effectiveness tests. There does not seem to be an interaction between those two fields, which could be of significant value for each other. For instance, several game developers have in the user requirements analyses and during personal talks stressed the added value of having an effectiveness label as a sort of selling strategy for their products. Independent academic researchers are the only credible source to provide them these labels. Moreover, academics consist of a lot of knowledge regarding motivation and learning theories, but do not have the creativity game developers have to translate this into engaging game mechanics. I believe that with this Ph.D. project we have

aimed and succeeded to bridge a gap between the academic field and the industry, by developing a procedure that attempts to find a balance between research rigor and practical limitations. We have also aimed to bridge this gap by listening to the needs of the several stakeholders involved in the DGBL field and coupling results back to those stakeholders. I believe that in the future more efforts should be devoted in bridging the gap between research and development in order to optimize the quality of the products. Not only by testing effectiveness, but also by sharing valuable theoretical insights we as academics have regarding development. This can only be achieved by providing these insights in a 'light' and easy readable jargon-free manner. We as researchers, should all aim to put some effort in sharing our results in this way, besides publishing in scientific journals, which mainly target an academic audience who are typically not developers.

Reference List

A

Adams, D. M., & Clark, D. B. (2014). Integrating self-explanation functionality into a complex game environment: Keeping gaming in motion. *Computers & Education, 73*, 149-159.

Adkins, S. (2015). The 2014-2019 global edugame market: Ambient insight. Research report retrieved from: http://www.ambientinsight.com/Resources/Documents/AmbientInsight_2014_2019_Global_Edugame_Market_Whitepaper.pdf

All, Nuñez Castellar, E. P., & Van Looy, J. (2014). Measuring Effectiveness in Digital Game-Based Learning: A Methodological Review. *International Journal of Serious Games, 1*(1).

Amory, Alan. 2010. "Learning to play games or playing games to learn? A health education case study with Soweto teenagers." Australasian Journal of Educational Technology 26(6):810-829.

Anastasi, A. (1961). *Differential psychology: Individual and group differences in behavior*: London, UK: Macmillan.

Ang, C. S., & Zaphiris, P. (2005). Developing Enjoyable Second Language Learning Software Tools: A Computer Game Paradigm (pp. 1-21). In C. S. Ang & P. Zaphiris (Eds.), *User-Centered Computer Aided Language Learning*. Hershey, PA: Information Science Publishing.

B

Baranowski, T., Buday, R., Thompson, D. I., & Baranowski, J. (2008). Playing for real: video games and stories for health-related behavior change. *American journal of preventive medicine, 34*(1), 74.

Bandura, A. (1986). Social foundations of thought and action. *Englewood Cliffs, NJ*.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology, 52*(1), 1-26.

Backlund, P., & Hendrix, M. (2013). *Educational Games-Are They Worth the Effort? a Literature Survey of the Effectiveness of Serious Games.* Paper presented at the 5th International Conference on. Games and Virtual Worlds for Serious Applications (VS-GAMES).

Baker, S. E., & Edwards, R. (2012). How many qualitative interviews is enough? National Centre for research Methods. Review paper retrieved from: http://eprints.brighton.ac.uk/11632/1/how_many_interviews.pdf

Bakker, M., van den Heuvel-Panhuizen, M., & Robitzsch, A. (2015). Effects of playing mathematics computer games on primary school students' multiplicative reasoning ability. *Contemporary Educational Psychology, 40*, 55-71.

Baranowski, T., Buday, R., Thompson, D. I., & Baranowski, J. (2008). Playing for real: video games and stories for health-related behavior change. *American journal of preventive medicine, 34*(1), 74.

Bardon, T., & Josserand, E. (2009). Digital game based learning: beyond pedagogical motivations. *Education and Training*.

Barzilai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. *Computers & Education, 70*, 65-79.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: an overview. *Advances in Human-Computer Interaction, 2013*, 1.

Boud, D., Keogh, R., & Walker, D. (2013). *Reflection: Turning experience into learning*. New York, NY: Routledge.

Boulianne, S. (2009). Does Internet use affect engagement? A meta-analysis of research. *Political communication, 26*(2), 193-211.

Bourgonjon, J., De Grove, F., De Smet, C., Van Looy, J., Soetaert, R., & Valcke, M. (2013). Acceptance of game-based learning by secondary school teachers. *Computers & Education, 67*, 21-35.

Boyle, E., Connolly, T. M., & Hainey, T. (2011). The role of psychology in understanding the impact of computer games. *Entertainment Computing, 2*(2), 69-74.

Braver, Mary W and Sanford L Braver. 1988. "Statistical treatment of the Solomon four-group design: A meta-analytic approach." Psychological Bulletin 104(1):150.

Breuer, J. S., & Bente, G. (2010). Why so serious? On the relation of serious games and learning. *Eludamos. Journal for Computer Game Culture, 4*(1), 7-24.

Brook, R. H., & Lohr, K. N. (1991). *Efficacy, Effectiveness, Variations, and Quality*: RAND Corporation.

Brom, C., Preuss, M., & Klement, D. (2011). Are educational computer micro-games engaging and effective for knowledge acquisition at high-schools? A quasi-experimental study. *Computers & Education, 57*(3), 1971-1988.

Brom, C., Šisler, V., Buchtová, M., Klement, D., & Levčík, D. (2012). Turning high-schools into laboratories? lessons learnt from studies of instructional effectiveness of digital games in the curricular schooling system *E-Learning and Games for Training, Education, Health and Sports* (pp. 41-53): Springer.

C

Cagiltay, N. E., Ozcelik, E., & Ozcelik, N. S. (2015). The effect of competition on learning in games. *Computers & Education*.

Calder, J. (2013). *Programme evaluation and quality: A comprehensive guide to setting up an evaluation system*: Oxon: Routledge.

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*: Houghton Mifflin Boston.

Chen, C.-H., & Law, V. (2016). Scaffolding individual and collaborative game-based learning in learning performance and intrinsic motivation. *Computers in Human Behavior, 55*, 1201-1212.

Chiang, Y.-T., Sunny, S., Chao-Yang, C., & Liu, E. Z.-F. (2011). Exploring online game players' flow experiences and positive affect. *TOJET: The Turkish Online Journal of Educational Technology, 10*(1).

Chiu, Y. h., Kao, C. w., & Reynolds, B. L. (2012). The relative effectiveness of digital game-based learning types in English as a foreign language setting: A meta-analysis. *British Journal of Educational Technology, 43*(4), E104-E107.

Clancey, W. J. (1991). Situated cognition: Stepping out of representational flatland. *AI Communications The European Journal on Artificial Intelligence, 4*(2/3), 109-112.

Clark, R.E. (1985). Confounding in educational computing research. *Journal of educational computing research, 1*(2), 137-148.

Clark, D (2007). Games and e-learning. *Caspian learning*. White paper retrieved from: http://www.itu.dk/~pgu/speciale/Whtp_caspian_games%201.1.pdf

Clark., R.E. (2007). Learning from serious games? Arguments, evidence, and research suggestions. *Educational technology, 47*(3), 56-59.

Clark, Tanner-Smith, E. E., & Killingsworth, S. S. (2015). Digital Games, Design, and Learning A Systematic Review and Meta-Analysis. *Review of Educational Research*, 0034654315582065.

Clark, D. (2007). Games, motivation & learning. *Caspian learning*.

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and* Development, 42(2), 21-29.

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. Self processes and development. The Minnesota symposia on child psychologyVol. 23., (pp. 43-77).

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. (2012). A systematic literature review of empirical evidence on computer games and serious games. Computers & Education, *59*(2), 661-686.

Connolly (2014). *Psychology, Pedagogy, and Assessment in Serious Games*: IGI Global

Cook, T. D. (1975). Sesame street revisited: Russell Sage Foundation.

Corsi, T. M., Boyson, S., Verbraeck, A., VAN HOUTEN, S.-P., Han, C., & MacDonald, J. R. (2006). The real-time global supply chain game: New educational tool for developing supply chain management professionals. Transportation Journal, 61-73.

Crawford, J., Stewart, L., & Moore, J. (1989). Demonstration of savings on the AVLT and development of a parallel form. *Journal of Clinical and Experimental Neuropsychology, 11*(6), 975-981.
Crookall, D. (2014). Engaging (in) Gameplay and (in) Debriefing. *Simulation & Gaming, 45*(4-5), 416-427.
Cruz, E. M. C., Cruz, J. A. V., Ruiz, J. G. R., David, L., & Hernández, H. (2015). Video Games in Teaching-Learning Processes: A Brief Review. *International Journal of Secondary Education, 2*(6), 102.
Csikszentmihalyi, M. (1990). Flow: The Psychology of Optimal Experience. New York, NY.
Csikszentmihalyi, M. (2013). Flow: The Psychology of Optimal Experience, 1990, Ney York. *NY Harper & Row*.

D

d'Ydewalle, G., & Van de Poel, M. (1999). Incidental foreign-language acquisition by children watching subtitled television programs. *Journal of Psycholinguistic Research, 28*(3), 227-244.
Daley, A. J. (2009). Can exergaming contribute to improving physical activity levels and health outcomes in children? *Pediatrics, 124*(2), 763-771.
De Freitas, S. (2006). Learning in immersive worlds. London: Joint Information Systems Committee.
De Grove, F., Bourgonjon, J., & Van Looy, J. (2012). Digital games in the classroom? A contextual approach to teachers' adoption intention of digital games in formal education. *Computers in Human Behavior, 28*(6), 2023-2033.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, *39*(1), 1-38.
DeSmet, A., Van Lippevelde, W., Van Ryckeghem, D., Compernolle, S., Bastiaensens, S., Poels, K., Vandebosch, H., et al. (2014). A systematic review and meta-analysis of serious digital games for healthy lifestyle promotion. ICA, 64th Annual conference, Papers. Presented at the ICA's 64th Annual conference: Communication and "the good life," International Communication Association (ICA).
Desurvire, H., Caplan, M., & Toth, J. A. (2004). Using heuristics to evaluate the playability of games. Paper presented at the CHI'04 extended abstracts on Human factors in computing systems.
Dickey, M. D. (2007). Game design and learning: A conjectural analysis of how massively multiple online role-playing games (MMORPGs) foster intrinsic motivation. *Educational Technology Research and Development, 55*(3), 253-273.
Dimitrov, D. M., & Rumrill, J., Phillip D. (2003). Pretest-posttest designs and measurement of change. *Work: A Journal of Prevention, Assessment and Rehabilitation, 20*(2), 159-165.
Dochy, Filip, Mien Segers and Michelle M Buehl. 1999. "The relation between assessment practices and outcomes of studies: The case of research on prior knowledge." Review of Educational Research 69(2):145-186.
Donovan, L., & Lead, P. (2012). The Use of Serious Games in the Corporate Sector. *A State of the Art Report. Learnovate Centre (December 2012)*.
Duffy, M. E. (2006). Handling missing data: a commonly encountered problem in quantitative research. *Clinical Nurse Specialist, 20*(6), 273-276.

E

Erhel, S., & Jamet, E. (2013). Digital game-based learning: Impact of instructions and feedback on motivation and learning effectiveness. *Computers & Education, 67*, 156-167.
Ermi, L., & Mäyrä, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. *Worlds in play: International perspectives on digital games research*, 37.
Estes, C. A. (2004). Promoting student-centered learning in experiential education. *Journal of Experiential Education, 27*(2), 141-160.

F

Fanning, R. M., & Gaba, D. M. (2007). The role of debriefing in simulation-based learning. *Simulation in healthcare, 2*(2), 115-125.
Federoff, M. A. (2002). Heuristics and usability guidelines for the creation and evaluation of fun in video games. Citeseer.

Field, A. (2009). Discovering statistics using SPSS. London: Sage publications
Fisher, R. A. (1934). Statistical methods for research workers. London: Paternoster Row.
Fisher, R. A. (1935). The design of experiments. New York, NY: Hafner Publishing Company.
Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., Mościcki, E. K., Schinke, S., Valentine, J. C., & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention science, 6*(3), 151-175.
Flick, U. (2009). *An introduction to qualitative research*: Sage.
Flick, U. (2011). *Introducing research methodology: A beginner's guide to doing a research project*: Sage.

G

Gagne, R. M. (1984). Learning outcomes and their effects: Useful categories of human performance. *American psychologist, 39*(4), 377.
Galarneau, L. L. (2005). Authentic learning experiences through play: Games, simulations and the construction of knowledge. *Paper presented at Digital Games Research Association (DiGRA), Vancouver, canada.*
Giannakos, M. N. (2013). Enjoy and learn with educational games: Examining factors affecting learning performance. *Computers & Education, 68*, 429-439.
Giessen, H. W. (2015). Serious Games Effects: An Overview. *Procedia-Social and Behavioral Sciences, 174*, 2240-2244.
Gunter, G. A., Kenny, R. F., & Vick, E. H. (2006). A case for a formal design paradigm for serious games. *The Journal of the International Digital Media and Arts Association, 3*(1), 93-105.
Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, Motivation, and Learning: A Research and Practice Model. Simulation & Gaming, 33(4), 441-467. doi: 10.1177/1046878102238607
Gerber, A. S., & Green, D. P. (2012). *Field Experiments. Design, Analysis and Interpretation.* New York, NY: W. W. Norton & Company.
Glaser, B. G., & Strauss, A. L. (2009). *The discovery of grounded theory: Strategies for qualitative research*: Transaction Books.
Gordon, T. J. (1994). The delphi method. *Futures research methodology, 2*.
Grabe, S., Ward, L. M., & Hyde, J. S. (2008). The role of the media in body image concerns among women: a meta-analysis of experimental and correlational studies. Psychological Bulletin, 134(3), 460.
Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: an experimental and individual difference investigation. *Journal of personality and social psychology, 52*(5), 890.

H

Hainey, T. (2010). *Using Games-Based Learning to Teach Requirements Collection and Analysis at Tertiary Education Level.* Retrieved from http://cis.uws.ac.uk/thomas.hainey/Final%20PhD%20Thesis%20Tom%20Hainey.pdf
Hainey, T., Connolly, T., Boyle, E., Azadegan, A., Wilson, A., Razak, A., & Gray, G. (2014). A Systematic Literature Review to Identify Empirical Evidence on the use of Games-Based Learning in Primary Education for Knowledge Acquisition and Content Understanding. *Proceedings of the 8th European Conference on Games Based Learning*, pp. 167-175.
Hauge, J. B., Boyle, E., Mayer, I., Nadolski, R., Riedel, J. C., Moreno-Ger, P., Bellotti, F., Lim, T., & Ritchie, J. (2014). Study Design and Data Gathering Guide for Serious Games' Evaluation. In T. M. Connolly, T. Hainey, E. Boyle, G. Baxter, & P. Moreno- Ger (Eds.), Psychology, Pedagogy, and Assessment in Serious Games. (pp. 394-419). Hershey, PA: IGI Global.
Hays, R. T. (2005). The effectiveness of instructional games: a literature review and discussion. Technical report retrieved from: http://faculty.uoit.ca/kapralos/csci5530/Papers/hays_instructionalGames.pdf
Hein, G. (1991). Constructivist learning theory. Institute for Inquiry. Available at:/http://www. exploratorium. edu/ifi/resources/constructivistlearning. htmlS.
Higgins, J. P., Green, S., & Collaboration, C. (2008). Cochrane handbook for systematic reviews of interventions (Vol. 5): Wiley Online Library.

Hoffman, D. L., & Novak, T. P. (2009). Flow online: lessons learned and future prospects. Journal of Interactive Marketing, 23(1), 23-34.

Huang, W.-H., Huang, W.-Y., & Tschopp, J. (2010). Sustaining iterative game playing processes in DGBL: The relationship between motivational processing and outcome processing. *Computers & Education, 55*(2), 789-797.

Hunsley, J., Elliott, K., & Therrien, Z. (2014). The efficacy and effectiveness of psychological treatments for mood, anxiety, and related disorders. *Canadian Psychology/Psychologie Canadienne, 55*(3), 161.

Hutchinson, L. (1999). Evaluating and researching the effectiveness of educational interventions. *BMJ: British Medical Journal, 318*(7193), 1267.

Hwa Hsu, S., Lee, F.-L., & Wu, M.-C. (2005). Designing action games for appealing to buyers. *CyberPsychology & Behavior, 8*(6), 585-591.

Hwang, G. J., & Wu, P. H. (2012). Advancements and trends in digital game-based learning research: a review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology, 43*(1), E6-E10.

J

Jarvin, L. (2015). Edutainment, Games, and the Future of Education in a Digital World. *New directions for child and adolescent development, 2015*(147), 33-40.

Joo, Y. J., Lim, K. Y., & Park, S. Y. (2011). Investigating the structural relationships among organisational support, learning flow, learners' satisfaction and learning transfer in corporate e-learning. *British Journal of Educational Technology, 42*(6), 973-984.

Joy, E. H., & Garcia, F. E. (2000). Measuring learning effectiveness: A new look at no-significant-difference findings. *Journal of Asynchronous Learning Networks, 4*(1), 33-39.

Juul, J. The game, the player, the world: Looking for a heart of gameness. In *Level up: Digital games research conference proceedings, 2003* (Vol. 120, pp. 121): Utrecht Univ., Utrecht, Holland

K

Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. *Handbook of research on effective electronic gaming in education, 1*, 1-32.

Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of instructional development, 10*(3), 2-10.

Keller, J. M. (2010). Motivational design for learning and performance: Springer.

Kharrazi, H., Lu, A. S., Gharghabi, F., & Coleman, W. (2012). A Scoping Review of Health Game Research: Past, Present, and Future. Games *for Health Journal, 1*(2), 153-164. doi: 10.1089/g4h.2012.0011

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

Kirriemuir, J., & McFarlane, A. (2004). Literature review in games and learning. A Report for NESTA Futurelab. http://www.futurelab.org.uk/resources/documents/lit_reviews/ Games_Review.pdf

Knapp, T. R and Schafer.W.D. (2009) From Gain Score t to ANCOVA F (and vice versa).Practical Assessment, *Research & Evaluation,14*(6):1-7.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development* (Vol. 1): Prentice-Hall Englewood Cliffs, NJ.

Korteling, J. E., Helsdingen, A., Sluimer, R., Emmerik, M. L., & Kappé, B. (2011). *Transfer of Gaming: transfer of training in serious gaming*: TNO innovation for life.

Korteling, J. E., Helsdingen, A., Sluimer, R., Emmerik, M. L., & Kappé, B. (2011). *Transfer of Gaming: transfer of training in serious gaming*: TNO innovation for life.

Kozma, R. B. (1994). Will Media Influence Learning? Reframing the Debate. . *Educational Technology Research and Development, 42*(2), 7-19.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of applied psychology, 78*(2), 311.

Kretschmann, R. (2012). Digital Sport-Management Games and Their Contribution to Prospective Sport-Managers' Competence Development. *Advances in Physical Education, 2*(4), 179-186.
Kriz, W. C. (2008). A systemic-constructivist approach to the facilitation and debriefing of simulations and games. *Simulation & Gaming,* 41, 663-680.
Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models. Boston, MA: McGraw-Hill Irwin.

L

Ladley, P. (2010). Games based situated learning: Games-ED whole class games and learning outcomes. *London, England: The Pixel Foundation Ltd. Retrieved fromhttp://www. pixelfountain. co. uk/download/Games-Based-Situated-Learning-v1. pdf*.
Lazzaro, N., & Keeker, K. (2004). *What's my method?: a game show on games.* Paper presented at the CHI'04 Extended Abstracts on Human Factors in Computing Systems.
Leary, M. R. (1995). *Introduction to behavioral research methods*: Brooks/Cole Pacific Grove, CA.
Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*(4), 612-625.
Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764-766.
Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement*.
Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley & Sons.

M

Maguire, M. (2003). *The use of focus groups for user requirement analysis*: Taylor & Francis, London.
Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction*. *Cognitive science, 5*(4), 333-369.
Mayer, R. (2011). Multimedia Learning and Games. In S. Tobias (Ed.), *Computer games and instruction* (pp. 281-306). Charlotte, NC: Information Age Publishing.
Mayer, I. (2012). Towards a Comprehensive Methodology for the Research and Evaluation of Serious Games. *Procedia Computer Science, 15*, 233-247.
Mayer, I., Bekebrede, G., Warmelink, H., & Zhou, Q. (2013). A Brief Methodology for Researching and Evaluating Serious Games and Game-Based Learning. In T. C. T. H. E. B. G. B. P. Moreno-Ger (Ed.), *Psychology, Pedagogy, and Assessment in Serious Games*: ICI Global.
Mayer, I., Bekebrede, G., Harteveld, C., Warmelink, H., Zhou, Q., Ruijven, T., Lo, J., Kortmann, R., & Wenzler, I. (2014). The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology, 45*(3), 502-527.
McCambridge, J., Butor-Bhavsar, K., Witton, J., & Elbourne, D. (2011). Can research assessments themselves cause bias in behaviour change trials? A systematic review of evidence from Solomon 4-group studies. *PLoS One, 6*(10), e25223.
McClelland, G. H. (2000). Nasty data. (pp. 393-411) In Reis, H.T. & Judd, C.M. (eds.) *Handbook of research methods in social psychology*. Cambridge, UK: Cambridge University Press.
Moser, C., Fuchsberger, V., & Tscheligi, M. (2012). Rapid assessment of game experiences in public settings. In *Proceedings of the 4th International Conference on Fun and Games* (pp. 73-82): ACM.
Michael, D. R., & Chen, S. L. (2005). *Serious games: Games that educate, train, and inform*: Muska & Lipman/Premier-Trade.
Michaud, L., Alvarez, J., Alvarez, V., & Djaouti, D. (2012). Serious Games: Enjeux, offre et marché. In idate (Ed.).
Miller, D. J., & Robertson, D. P. (2010). Using a games console in the primary classroom: Effects of 'Brain Training' programme on computation and self-esteem. *British Journal of Educational Technology, 41*(2), 242-255. doi: 10.1111/j.1467-8535.2008.00918.x

Miller, D. J., & Robertson, D. P. (2011). Educational benefits of using game consoles in a primary classroom: A randomised controlled trial. *British Journal of Educational Technology, 42*(5), 850-864. doi: 10.1111/j.1467-8535.2010.01114.x

McQuail, D. (2010). *McQuail's mass communication theory*: Sage publications.

Nabi, R. L., & Oliver, M. B. (2009). *The SAGE handbook of media processes and effects*: Sage.

N

Neys, J., Van Looy, J., De Grove, F., & Jansz, J. (2012). *Poverty is not a game: behavioral changes and long term effects after playing PING.* Paper presented at the 13th annual conference on the International Speech Communication Association, Portland.

Newzoo. (2015a). Global report: US and China tak ehalf of $113BN games market in 2018. In Newzoo (Ed.): Newzoo. Retrieved from: https://newzoo.com/insights/articles/us-and-china-take-half-of-113bn-games-market-in-2018/

Newzoo. (2015b). Newzoo summer series #24: Belgian Games Market. Retrieved from: https://newzoo.com/insights/infographics/newzoo-summer-series-24-belgian-games-market/

Nicholson, S. (2012). *Completing the experience: Debriefing in experiential educational games.* Proceedings of The 3rd International Conference on Society and Information Technologies (pp. 117-121). Winter Garden, FL.

Nuñez Castellar, E., Van Looy, J., Szmalec, A., & De Marez, L. (2013). Improving arithmetic skills through gameplay: assessment of the effectiveness of an educational game in terms of cognitive and affective learning outcomes. *Information Sciences, 264*(April 2014), 19–31

Nussbaum, M., & Beserra, V. d. S. (2014). Educational Videogame Design. In *Proceedings of the 14th International Conference on Advanced learning technologies* (pp. 2-3). Athens, Greece: IEEE.

O

O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *The Cirriculum Journal, 16*(4), 455-474.

P

Pagulayan, R. J., Keeker, K., Wixon, D., Romero, R. L., & Fuller, T. (2002). User-centered design in games.

Pannese, L., Cassola, M., & Grassi, M. (2005). *Interaction with simulation tools: analysis of use cases.* Paper presented at the I-KNOW Conference.

Pannese, L., & Carlesi, M. (2007). Games and learning come together to maximise effectiveness: The challenge of bridging the gap. *British Journal of Educational Technology, 38*(3), 438-454.

Paras, B. (2005). Game, motivation, and effective learning: An integrated model for educational game design.

Parchman, S. W., Ellis, J. A., Christinaz, D., & Vogel, M. (2000). An Evaluation of Three Computer-Based Instructional Strategies in Basic Electricity and Electronics Training. . *Military Psychology, 12*(1), 73-87.

Pearce, L. J., & Field, A. P. (2016). The Impact of “Scary” TV and Film on Children's Internalizing Emotions: A Meta-Analysis. *Human Communication Research, 42*(1), 98-121.

Perse, E. M. (2001). *Media effects and society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pittenger, A., & Doering, A. (2010). Influence of motivational design on completion rates in online self-study pharmacy-content courses. *Distance Education, 31*(3), 275-293.

Pivec, M. (2007). Editorial: Play and learn: potentials of game-based learning. *British Journal of Educational Technology, 38*(3), 387-393.

Popper, K. (2000). Science: conjectures and refutations. *Readings in the Philosophy of Science: From Positivism to Postmodernism*, 9-13.

Prensky, M. (2001). *Digital game-based learning*. New York, NY: McGraw-Hill.

R

Randel, J. M., Morris, B. A., Wetzel, C. D., & Whitehill, B. V. (1992). The effectiveness of games for educational purposes: A review of recent research. *Simulation & Gaming, 23*(3), 261-276.

Ritterfeld , U., & Weber, R. (2006). Video games for entertainment and education. *Playing Video Games. Motives, Responses, and Consequences. Mahwah, NJ: Lawrence Erlbaum Associates*, 399-413.

Robertson, J. (2006). *Mastering The Requirements Process, 2/E*: Pearson Education India.

Rooney, P. (2012). A Theoretical Framework for Serious Game Design: Exploring Pedagogy, Play and Fidelity and their Implications for the Design Process. *International Journal of Game-Based Learning, 2*(4), 41-60.

Rubin, D. B. (1973). Matching to remove bias in observational studies. Biometrics, *29*(1), 159-183.

Ryan, R. M., & Deci, E. L. (2000a). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25*(1), 54-67.

Ryan, R. M., & Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist, 55*(1), 68.

S

Savery, J. R., & Duffy, T. M. (1995). Problem based learning: An instructional model and its constructivist framework. *Educational technology, 35*(5), 31-38.

Sawyer, B. and P. Smith, Taxonomy for Serious Games. Digitalmil, Inc& Serious Games Initiative/Univ. of Central Florida, RETRO Lab, 2008.

Scheffer, J. (2002). Dealing with missing data*. Research Letters in the Information and Mathematical Sciences 3*(1), 153–160.

Schønau-Fog, H., & Bjørner, T. (2012). "Sure, I Would Like to Continue" A Method for Mapping the Experience of Engagement in Video Games. *Bulletin of Science, Technology & Society, 32*(5), 405-412.

Serrano-Laguna, Á., Torrente, J., Manero, B., Blanco, Á. d., Blanca, Borro-Escribano, . . . Fernández-Manjón, B. (2013*). Learning Analytics and Educational Games: Lessons Learned from Practical Experience.* Paper presented at the Games and Learning Alliance Conference, Paris.

Short, H. (2014). A critical evaluation of the contribution of trust to effective Technology Enhanced Learning in the workplace: A literature review*. British Journal of Educational Technology, 45*(6), 1014-1022.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*: Oxford university press.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel psychology, 64*(2), 489-528.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction, 55*(2), 503-524.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel psychology, 64*(2), 489-528.

Smith, S. L., & Granados, A. D. (2009). Content patterns and effects surrounding sex-role stereotyping on television and film. *Media effects: Advances in theory and research, 3*, 342-361.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin, 46*(2), 137.

Song, H., Zmyslinski-Seelig, A., Kim, J., Drent, A., Victor, A., Omori, K., & Allen, M. (2014). Does Facebook make you lonely?: A meta analysis. *Computers in Human Behavior, 36*, 446-452.

Skadberg, Y. X., & Kimmel, J. R. (2004). Visitors' flow experience while browsing a Web site: its measurement, contributing factors and consequences. *Computers in Human Behavior, 20*(3), 403-422.

Stewart, J., Bleumers, L., Van Looy, J., Mariën, I., All, A., Schurmans, D., Willaert, K., De Grove, F., Jacobs, A., & Misuraca, G. (2013). The Potential of Digital Games for Empowerment and Social Inclusion of Groups at Risk of Social and Economic Exclusion: Evidence and Opportunity for Policy. In: Institute for Prospective and Technological Studies, Joint Research Centre.

Susi, T., Johannesson, M., & Backlund, P. (2007). Serious games: An overview. *Technical report retrieved from: http://www.diva-portal.org/smash/get/diva2:2416/FULLTEXT01.pdf*

Suddaby, R. (2006). From the editors: What grounded theory is not. *Academy of management journal, 49*(4), 633-642.
Strauss, A., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*: Sage Publications, Inc.
Sweetser, P., & Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE), 3*(3), 3-3.

T

Tannenbaum, S. I., & Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 55*(1), 231-245.
Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478.
Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation, 27*(2), 237-246.
Tsai, F.-H., Yu, K.-C., & Hsiao, H.-S. (2012). Exploring the Factors Influencing Learning Effectiveness in Digital Game-based Learning. *Educational Technology & Society, 15*(3), 240-250.

V

Vallerand, R. J., & Reid, G. (1984). On the causal effects of perceived competence on intrinsic motivation: A test of cognitive evaluation theory. *Journal of Sport Psychology, 6*(1), 94-102.
Valkenburg, P. M., Peter, J., & Walther, J. B. (2016). Media Effects: Theory and Research. *Annual review of psychology, 67*, 315-338.
Van Eck, R. (2015). Digital Game-Based Learning: Still Restless, After All These Years. *EDUCAUSE review, november/december 2015*, 13-28.
van Engelenburg, G. (1999). Statistical Analysis for the Solomon Four-Group Design. Research Report 99-06 retrieved from: http://files.eric.ed.gov/fulltext/ED435692.pdf
Van Der Meij, H., Leemkuil, H., & Li, J.-L. (2013). Does individual or collaborative self-debriefing better enhance learning from games? *Computers in Human Behavior, 29*(6), 2471-2479.
Van Looy, J., Núñez Castellar, E., Houttekier, E. (submitted, 2016, 9-12 November). Relative Enjoyment Scale for Primary School Children (RES-C): Development and Testing of Reliability, Validity and Sensitivity. Paper to be presented at 6th European Communication Conference: Mediated (Dis)Continuities: Contesting Pasts, Presents and Futures, Prague, Czechoslovakia.
Vrugte, J., Jong, T., Wouters, P., Vandercruysse, S., Elen, J., & Oostendorp, H. (2015). When a game supports prevocational math education but integrated reflection does not. *Journal of Computer Assisted Learning, 31*(5), pp. 462-480.

W

Wellman, R. J., Sugarman, D. B., DiFranza, J. R., & Winickoff, J. P. (2006). The extent to which tobacco marketing and tobacco use in films contribute to children's use of tobacco: a meta-analysis. Archives of pediatrics & adolescent medicine, 160(12), 1285-1296.
Wilkes, M., & Bligh, J. (1999). Evaluating educational interventions. *BMJ: British Medical Journal, 318*(7193), 1269.
Wimmer, R. D., & Dominick, J. R. (2013). Mass media research. Boston, MA: Cengage learning
Wouters, P., van der Spek, E., & Van Oostendorp, H. (2009). Current practices in serious game research: A review from a learning outcomes perspective. *Games-based learning advancements for multi-sensory human computer interfaces: techniques and effective practices*, 232-250.
Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van Der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249.

Y

Yang, Y.-T. C. (2012). Building virtual cities, inspiring intelligent citizens: Digital games for developing students' problem solving and learning motivation. *Computers & Education, 59*(2), 365-377. doi: 10.1016/j.compedu.2012.01.012

Yip, F. W. M., & Kwan, A. C. M. (2006). Online vocabulary games as a tool for teaching and learning English vocabulary. Educational Media International, 43(3), 233-249. doi: 10.1080/09523980600641445

Z

Zhang, D., Zhao, J. L., Zhou, L., & Nunamaker Jr, J. F. (2004). Can e-learning replace classroom learning? *Communications of the ACM, 47*(5), 75-79.

Zimmerman, B. J., & Schunk, D. H. (2003). Albert Bandura: The Man and his Contributions to Educational Psychology. In B. J. Zimmerman & D. H. Schunk (Eds.), *Educational psychology: One-hundred years of contributions.* . Mahwah, New Jersey: Lawrence Erlbaum Associates.

Appendix:
Overview of contents: Procedure V 2.0

In consultation with Ghent University, we have decided to not make the procedure publically available as it is intellectual property of Ghent University. Below, a detailed overview of the contents of the procedure can be found.