An imperfect union? Enacting an analytic and evaluative framework for digital games for language learning

Carolyn Blume,[1] Torben Schmidt[2] und Inke Schmidt[3]

Begriffe wie Gamification und Serious Games für computerbasierte Lernanwendungen einerseits, sowie die Einbindung von Unterhaltungsspielen in Wissens- und Kompetenzerwerbsprozesse andererseits stehen exemplarisch für disziplinenübergreifende Ansätze, durch die Nutzung von Spielmechaniken das Lernen zu optimieren. Auch für den Bereich des *Computer-Assisted Language Learnings* (*CALL*) ist dies festzustellen. Während die bisherige Forschung sich insbesondere mit theoretischen Aspekten einzelner Anwendungen befasste oder in empirischen Analysen sich den (fremdsprachen-)förderlichen Elementen bezüglich einzelner Kompetenzbereiche widmete, stellen Untersuchungen der Charakteristika der Programme sowie Ansätze zur Evaluierung ein Forschungsdesiderat dar. Der vorliegende Beitrag stellt die Prozesse der Entwicklung und Implementierung eines Test- und Prüfinstruments[4] in den Fokus, das Programme u.a. hinsichtlich ihrer lerntheoretischen Fundierung, ihrer didaktischen Merkmale, der genutzten Spielmechaniken sowie der multimedialen Gestaltung analysiert. Anschließend folgt die Vorstellung und Diskussion zentraler Ergebnissen der Analyse von 50 gegenwärtig erhältlichen Programmen. Der Beitrag schließt mit Betrachtungen zur Weiterentwicklung des Test- und Prüfinstruments sowie Empfehlungen zur Gestaltung zukünftiger Lernprogramme.

1. Introduction

Despite the fact that digital game-based language learning (DGBLL) has recently emerged as a substantial area of interest in educational research and practice, fueled by a range of pedagogical and pragmatic interests, a significant hindrance to the instructional use of applications results from the extensive range of games available, and the language educator's need to select games consistent with his or her pedagogical aims and the target population's developmental stages, academic skills, attitudinal biases, and available technological resources (cf. Burston 2003). In the absence of explicit information regarding applications' underlying didactic[5] models, educators and learners are given little guidance regarding the degree to

1 Korrespondenzadresse: Carolyn Blume, Leuphana Universität Lüneburg, Zukunftszentrum Lehrerbildung, Universitätsallee 1, 21335 Lüneburg, E-Mail: cblume@leuphana.de
2 Korrespondenzadresse: Prof. Dr. Torben Schmidt, Leuphana Universität Lüneburg, Institute of English Studies, Universitätsallee 1, 21335 Lüneburg, E-Mail: torben.schmidt@leuphana.de
3 Korrespondenzadresse: Inke Schmidt, FAU Erlangen-Nürnberg, Institut für Lern-Innovation, Dr.-Mack-Str. 77, 90762 Fürth, E-Mail: inke.schmidt@ili.fau.de
4 http://ludicall.de/sites/default/files/pruefbogen.pdf. Last accessed 04.09.2017.
5 "Didactic" is utilized throughout this article to describe the "design science" of determining "what to teach and how to teach" (Reusser 1996: 83).

which principles of foreign language learning, as delineated by Ellis (2005), inform the available tools. Moreover, while there has been relatively substantial research on commercial, off the shelf games (COTS) utilized for language learning purposes, there are few outcome-related analyses for dedicated applications. In addition to lingering questions regarding efficacy (Chiu, Kao & Reynolds 2012: E106), the complexity of identifying and evaluating appropriate applications for language learning make it challenging for educators to take full advantage of the range of available game-like programs with potentially significant affordances.

In order to support educators and researchers in appropriately selecting from among the many language learning games currently available, an evaluation framework that includes elements from the fields of foreign language learning theory, media pedagogy, computer-assisted language learning (CALL), digital game-based learning (DGBL), and game design was constructed. The purposes of this tool are to (1) provide guidance to potential users in selecting games that best meet their – or, their language learners' – needs and to (2) enable a systematic evaluation of available digital games for foreign language learning. Hence, this work finds itself in the "research-evaluation nexus" (Levy & Stockwell 2006: 41). It provides a set of descriptive criteria that enable a substantial number of applications to be analyzed with respect to a variety of pedagogical and multi-medial elements. While it does not attempt to determine the quality of the individual programs under review, it lays the foundation for such subsequent examinations.

This paper introduces the tool, and outlines the results of the initial analysis conducted using it, in order to illuminate the educational and gameplay-specific features of available DGBLL. Following a description of the existing research in evaluating CALL and DGBLL, a methodological overview of the evaluation instrument's conception, development, and implementation is presented. The article then presents an analysis of selected results. This data-based analysis facilitates a critical examination of the tool's efficacy and utility for the study of DGBLL.

2. Literature review

Whereas introspective evaluation of CALL programs dates back to the 1980s, research in DGBLL has focused on efficacy evaluation rather than pre-use analysis (cf. Hwang & Wu 2012; McMurry, West, Rich, Williams, Anderson & Hartshorn 2016). As a result, there is, on the one hand, a body of literature that supports methodological and pedagogical analyses of CALL that does not focus on game-based features (cf. Leakey 2011). On the other hand, there are numerous studies examining the acquisition of language knowledge through the use of digital game-

based applications (cf. Chiu et al. 2012) that do not comprehensively analyze language-related features of existing DGBLL.

2.1 Evaluation in the field of CALL

While the field of CALL was still in its infancy, Hubbard (1987) describes the challenges of evaluating it, the existing tools that attempt it, as well as the shortcomings of these approaches. He elucidates the need to examine applications in the context of specific language learning approaches, learner strategies, and syllabus types, with the focus on potential adoption by educators and related stakeholders. Subsequent revision of his proposed framework elaborates on these core components while acknowledging the continuing complexities associated with analyzing CALL applications (Hubbard 2006).

Like Hubbard, Chapelle (2001: 55) examines the construct of learner fit without delving into specific game-based features. In advocating an evaluation of CALL informed by theories of instructed SLA regarding attentional manipulation, for example, Chapelle (2001: 49) examines factors that are integral elements of digital games, including issues of pacing, multimodality, support, and control. While critical to CALL and digital games alike, she does not elaborate on the ways by which the latter structure these elements to support the former. She identifies a total of six areas that are relevant to CALL evaluation. In addition to learner fit, these include the importance of language learning potential, meaning focus, authenticity, positive impact, and practicality (Chapelle 2001: 55).

Tomlinson (2014) addresses both the content-related and procedural criteria necessary for an examination of computer-based language learning materials. Moreover, he distinguishes between analysis and evaluation, and the various types of queries required for each. While he emphasizes the importance of articulating the principles underlying both descriptive and analytic items, he addresses CALL only in the context of guidelines for developing self-access materials derived from what he terms "universal principles", simultaneously acknowledging that different developers will subscribe to different principles (Tomlinson 2010:73).

Despite these various efforts in the area of CALL, Roche (2003: 94) points out the inadequacy of existing formative evaluative processes that, lacking theoretical underpinnings, unsystematically examine learning processes in the pursuit of specific language learning aims. Existing evaluatory frameworks for CALL, according to McMurry et al. (2016), furthermore inadequately address explicit evaluation-relevant processes that would strengthen their utility. Neither Hubbard's nor Chapelle's framework, they argue as an example, engage in a meta-evaluation of the evaluation framework. They further suggest that the range of evaluands needs

to be broadened to include, for example, autonomous language learning activities. These criticisms echo the concerns of those who focus on DGBL (without the language specification) and who cite the lack of empirically grounded and validated models as a hindrance to meaningful analysis (Mayer, Bekebrede, Harteveld, Warmelink, Zhou, van Ruijven, Lo, Kortmann & Wenzler 2014: 509). Reeder, Heift, Roche, Tabyanian, Schlickau & Gölz (2004: 256) do not focus specifically on games, but allude to them in their description of more recent generations of CALL software, positing that existing evaluative criteria are inadequate for these newer products designed around contemporary understandings of both language pedagogy and technical design.

2.2 Evaluation of DGB(L)L

Although evaluative tools exist for DGBL, their emphasis differs from those available for CALL in both intent and focus. The construction of evaluation tools for DGBL has focused primarily on identifying relevant design principles (cf. Moreno-Ger, Burgos, Martínez-Ortiz, Sierra & Fernández-Manjón 2008) and on outcome-oriented instruments designed to measure efficacy (cf. Mayer et al. 2014: 503).

There are a few models that attempt to provide an evaluative framework for game-based learning (cf. Freitas & Oliver 2006; Carvalho, Bellotti, Berta, Gloria, Sedano, Hauge, Hu & Rauterberg 2015). Designed to be applicable for a variety of subjects and levels, these tools do not focus specifically on the unique features of DGBLL. Sykes & Reinhardt (2013: 150-152) provide a framework for evaluating commercial off-the-shelf games COTS repurposed for language learning. Given that their focus is on commercial games, underlying language learning theories and principles are not examined. Moreover, the open-ended formulation of the questions lends itself to the in-depth examination of one, but not a range of, potential tools.

Without developing a specific framework, Hubbard (1991: 221) early on identifies two critical issues in analyzing computer games for language learning. Grappling with the question as to what constitutes a game, he advances the notion that a program could be considered a game on the basis of students' intrinsic desire to engage with it. The motivational benefits ascribed to gaming have, in subsequent years, led to substantial theoretical and empirical analysis designed to illuminate this field (cf. Henry 2013; Schmidt, Schmidt & Schmidt 2016). While each of these approaches contributes to a better understanding of "ludic engagement" (Cornillie, Thorne & Desmet 2012: 243), no unifying model exists.

Hubbard's second question is whether the activity in question truly promotes language learning. It is not enough for an activity to be engaging; a "ludic methodology" must incorporate meaningful learning (Lombardi 2012). While issues of learner and teacher fit, as well as appropriateness judgments, play as much of a role as they do in evaluating non-game-based CALL programs, Hubbard emphasizes that the value of language learning activity presented in a game-like manner requires its own analysis of the language skill, content, or competence pursued in the game context.

The evaluation tool presented herein thus seeks to provide guidance as regards these two criteria. The questions in focus here assess the presence of game-like features that, according to existing literature from game studies, facilitate engagement (Arnab, Lim, Carvalho, Bellotti, Freitas, Louchart, Suttie, Berta & Gloria 2015: 397). Moreover, the ways in which the applications facilitate language learning are analyzed in terms of language learning theories and instructional approaches. The tool seeks to provide information about preponderant elements found in these applications. Such a foundation is necessary prior to assessing the quality of specific programs or genres, which has, as of yet, only rarely taken place, although there are exceptions (cf. Feist 2008; Kerres & Bormann 2009; Anyaegbu, Ting & Li 2012; Vesselinov & Grego 2012; Kober, Paland-Riedmüller & Hafner 2015, Vesselinov & Grego 2016a, 2016b; Chalak & Ahmadi 2017). Before the tool itself and the evaluation of games using it are addressed, the research design is described.

3. Research design

This section summarizes both phases of the study, namely the design of a tool for DGBLL evaluation and the application of said tool to existing programs. First, it summarizes the criteria by which tested applications were defined and identified. It then describes the constructs from existing principles and frameworks for evaluating CALL and DGBL that have been adapted to develop a comprehensive tool for DGBLL evaluation. Given space limitations, the focus in both describing the evaluation tool and the empirical results is on the defining characteristics for DGBLL, with a focus on foreign language learning theories and didactic and medial interactivity. The purpose of applying the tool in this study is not to evaluate individual products, but to examine the utility of the tool in use. Moreover, it facilitates the examination of widely implemented features of existing applications, in order to provide some initial conclusions about the characteristics of these programs in light of CALL and DGBL theory.

3.1 Application definition

In selecting applications, the researchers included programs, platforms, and applications that either define themselves as games or that use gamified elements to promote language learning. In the DGBL literature, a variety of terms are used. While commercial games (COTs) are generally considered to be distinct from serious games, games for learning, and synthetic immersive environments (cf. Breuer & Bente 2010; Sykes 2013), virtual worlds may be either commercial products adapted for learning or purpose-built environments (Peachey & Childs 2011: 2). The "use of game design elements in non-game contexts", known as gamification (Deterding, Dixon, Khaled & Nacke 2011: 10), may refer to either commercial or educational applications. While there is thus no consensus regarding what constitutes a "game" (Crookall 2010: 904), or even what to call DGBLL applications (Cornillie et al. 2012: 246), including all of these items assumes that products that identify themselves as games could be evaluated within those parameters and in light of what Deterding et al. (2011: 13) describe as the "socio-cultural trend of ludification."

Recognizing that the boundaries between the programs and their ancillary chats, blogs, discussion forums, walkthroughs, and websites are often nebulous (Karppi & Sotamaa 2012: 414), analysis of the applications includes examinations of these elements as well. This is what, to varying degrees, Consalvo (2007: 21) refers to as "paratexts", Salen & Zimmerman (2003: 431) as the "meta-game", and Gee & Hayes (2012: 130) as "Game". By adopting this approach, a wide net was cast to include a range of learning objects.

3.2 Application selection

The first step involved creating a database of relevant programs. In addition to engaging in Google-based searches, the researchers identified titles based on reviews from generic (i.e. "App store") and specialized DGBL (i.e. www.dji.de/; www.gamesforchange.org/learn/game-databases) websites. Also included were the offerings of language education institutes such as the *Goethe-Institut* and British Council. Despite the lack of verifiability regarding usage statistics, a program's popularity, based on downloads (i.e. www.appannie.com), was considered as one criterion for inclusion. In this initial step, a list of approximately 150 applications was compiled.

It was the goal to survey a range of applications, games, and programs, in order to examine a cross-section of tools on a variety of platforms, and therefore purposive sampling (cf. Teddlie & Yu 2007: 80-81) was utilized. At the same time, it

became clear that there is an uneven distribution of language learning applications among platforms, monetization models, and languages. For example, the most popular delivery method continues to be via browser. Based on these criteria of popularity, language-learning focus, and availability to the public, 50 applications that represented all major platforms, monetization models, game types, and developers were selected.

3.3 Evaluation tool development

Prior to analyzing the existing digital games for language learning, an appropriate evaluative tool reflecting, as per Hubbard (2006: 5), both methodological frameworks and aspects of instructed SLA theory was constructed in a four-phase, recursive process, with special attention given to Chapelle's principles for designing meaningful CALL evaluation (see 2.1 above).

Whenever possible, complex pedagogical or design principles have been operationalized as quantifiable variables. Thus, Hubbard's (1987: 236) descriptor, "provides comprehensible input at a level just beyond that currently acquired by the learner", is concretized in items that analyze the ability of the given program to adapt to a learner's level. However, the complexity of DGBLL cannot be fully captured by unitary measures. Where possible, scales of multiple items generate numeric or averaged values. In cases where this is too reductive, narrative descriptions describe the unique elements of the various applications.

These aspects serve as components of a comprehensive tool that ultimately considers 25 separate elements of DGBLL with 80 items. Complexity can be accounted for by the ability to assign multiple codes and through the provision of open-ended response fields. In some instances, Likert scales with five-point response items are used to evaluate constructs, especially when subjective analysis seems called for, i.e. in determining the degree of immersion or the quality of multimodal elements in individual applications. In this way, both quantitative and qualitative findings could be incorporated into what Hubbard terms a "principled checklist" (Hubbard 2006: 6). The final tool incorporates five categories, three of which are described more fully below. Given that the reported results do not address items in the two remaining categories (background information and user experience), these will be mentioned only briefly in terms of their function in the evaluative tool.

3.3.1 Didactic analysis

Bopp (2006: 10) points out that analyses of DGBL must examine the explicit and implicit learning goals, content, and didactic methods of the application. These items, alongside an examination of the activities, exercises and tasks that contribute to a focus on form, meaning, or both (cf. Chapelle 2001) and feedback types and timing (cf. Shute 2008; Conati & Manske 2009) facilitate an analysis of the underlying pedagogical theory and instructional model (cf. Bopp 2006: 21). While a listing of the content, competencies, and activities of the applications was undertaken before being subsequently coded and categorized, other aspects of the didactic analysis require a more integrated approach. Determining, for example, whether an application is largely behaviorist, cognitivist, or constructivist requires an analysis of multiple items that must take into account not only what is presented, but the ways in which it is presented.

Table 1: Didactic analysis

Didactic analysis	**Subcategory questions**
Proficiency level	- To what degree does the given proficiency level reflect the actual level?
Pre-test	- Is a pre-test available or required? - Do the results of the pre-test affect the content, sequence, or presentation of material within the application?
Competencies	- Which competencies and skills are addressed?
Content	- What is the content and what themes are found in the application? - In what ways is there an integration of language and content (CLIL)?
Instructional approach	- Is the language conveyed primarily implicitly or explicitly?
Additional support	- What opportunities are there for teachers/parents to adapt content or skill levels; see results; obtain off-line materials; engage simultaneously in the application?
Learning theories	- To what degree are behaviorist, cognitivist, and/or constructivist elements present in the application?
Activity/exercise/ task types	- To what degree are closed, semi-open, or open-ended activities integrated in the application? - To what degree are there elements that focus on form and/or on meaning?
Feedback	- What forms of feedback are utilized in the application? - To what extent can the user modify the feedback options?
Quality	- Are errors in content, explanations, language use, or feedback present?

3.3.2 Didactic interactivity

The concept of didactic interactivity draws on the definition of Strzebkowski & Kleeberg (2002: 232), who refer to it as those active interactions in educational software that directly support cognitive processes. Strzebkowski and Kleeberg differentiate between interactivity of design elements (*Steuerungsinteraktivität*) and didactic interactivity (*didaktische Interaktivität*), giving as examples of the latter the ability to influence animations, models, and simulations; input complex responses to complex queries; modify the content or progression; create new multimedia objects; utilize an electronic notebook; take advantage of adaptive feedback and help (ibid.). Macro- and micro-adaptations based on the proficiency level of the user and customized with the appropriate amount of scaffolding an individual learner requires create personalization (cf. Leutner 2002), evaluated by queries that examine how the application changes based on the apparent proficiency of the user.

Table 2: Didactic interactivity

Subcategory	**Question(s)**
Customization	- How does the application adapt to the user, either through active selection or passively through user behavior? - How does the application adapt to the user in terms of content, level of difficulty, learning style, or other characteristics?
Personal profile	- How is data about the user utilized to adapt the application in terms of content, presentation, or gameplay?
Scaffolding	- What is the nature of support to users in terms of content and gameplay?
User-created material	- To what extent does the application facilitate the integration of user-created material?

While the notion of what constitutes adequate adaptivity is based on a subjective determination, multi-layered analyses of how an application accommodates the users generates a descriptive measure of the application. Thus, while one application might incorporate a pre-test that affects the subsequent level of the material to be learned, another might automatically adapt to the user's behavior to alter the content presented to the learner.

3.3.3 Game-based characteristics

In this category, the examination of game-based features is brought into focus. While Sicart (2008) defines game mechanics as "methods invoked by agents for

interacting with the game world", giving as examples "climbing, jumping, stabbing and shooting", Arnab et al. (2015: 397) list both the aforementioned as well as more concrete and bounded items as tokens, rewards, and goods, and furthermore game*play* mechanics, such as levels, competition, and infinite play. The questions in this category accommodate both definitions.

Queries were generated to determine whether learners have the opportunity or are required to utilize critical thinking skills, based on a revised Bloom's Taxonomy that adapts the original hierarchical model of cognitive activities to reflect their process-oriented nature (cf. Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths & Wittrock 2001); whether the application's language learning activities are wholly or partially embedded in a narrative environment (cf. Ritterfeld & Weber 2006); and the nature of the generated outcomes. This is further delineated into subqueries, i.e. in the case of higher-order thinking skills, it is necessary to assess the degree to which activities potentially facilitate critical thinking as previously defined; to distinguish whether they are thoroughgoing, partially present, or absent; and to determine the degree to which the user is required, encouraged, or enabled to engage in these activities. A similar process is then necessary for the other elements, i.e. regarding the degree to which a narrative exists and the extent to which the narrative is integral to the language learning activities and goals. Finally, the rewards and goals require definition, and an analysis of their relevance towards the thinking skills, narrative, and language learning activities and goals is necessary (cf. Arnab et al. 2015).

Table 3: Game-based characteristics

Subcategory	**Question(s)**
Game genres	- Is there a narrative (story) underlying the application? - What game genre does the application belong to?
Game mechanics	- What game mechanics are present?
Game elements	- What game elements are present? - What do the users "do" in the application?
Social mechanics	- What social mechanics are present? - What are the functions of the existing social mechanics?
Real-world connections	- To what extent does the application make connections to the real world?

Evidence that both narrative and interaction, and especially their interaction, facilitate language learning informs both sociocultural theories of language learning and analysis of DGBL, highlighting the significance of social mechanics in DGBLL (cf. Reinders & Wattana 2011). The evaluation thus further includes items that address the question posed by Consalvo (2011: 188): how do games implement social interactions into their gameplay, to what purpose?

3.3.4 Background and user experience

While data regarding program specifications (version, platforms, monetization model) and user experience are gathered, these are not discussed further in this article. Although these categories are important, they are not examined for two different reasons. On the one hand, the collected background information, with one exception, does not provide further information that helps evaluate the pedagogic usefulness of the applications. On the other hand, in its role in shaping the user's options for interacting with the interface and the game world (Saltzman 2000: 261), user experience is closely related to the adaptivity required for effecttive didactic interactivity (Hochleitner, Hochleitner, Graf & Tscheligi 2015: 199), so that there exists overlap in these categories. It was thus decided to avoid repetition by focusing on those elements most closely related to issues of DGBLL (i.e. didactic analysis, didactic interactivity, and game-based mechanics).

Given the complexity of analyzing the effective utilization of multimedia elements for language learning, a substantial number of individual items in the remaining three categories – didactics analysis, didactic interactivity, and game-based characteristics – are addressed via Likert scales. This allows for analysis of the topic's breadth and the evaluative stance necessary for a meaningful analysis.

3.4 Implementation processes

In the second phase of development, the survey's reliability was tested. While the large number of qualitative items in the survey made a global correlation of inter-rater reliability impractical, collaborative coding of various game elements facilitated consistent analysis. To more adequately address the transdisciplinary nature of the tool, a manual was designed to accompany the checklist, elaborating on the underlying constructs and providing selected examples. A recursive dialogue further clarified queries and responses to heighten consistency among raters. In the final pilot phase, the testers completed game tests together, subsequently acting as "critical player-theorists", as described by Aarseth (2014: 181), as a form of action research (cf. Karppi & Sotamaa 2012).

Hubbard (2006: 1) notes that one obstacle to evaluating early CALL software arises from the fact that these applications do not generally enable the evaluator to "skim" the program, but rather, necessitate the application to be tested in its entirety, a challenging task given the complexity and number of available options in these programs. This is even more so the case with digital games, with their multiply branching options (Burston 2003: 35) Thus, the selected digital applications were "played" several times, in order to reveal the affordances the program offered for various learning pathways. This approach simultaneously offered the

opportunity for multiple testers to analyze each application and thus enhance inter-rater reliability.

In the following section, selected results derived from the implementation of this evaluation tool for these 50 applications are presented. While substantially more data was collected, the focus on these aspects allows for initial conclusions to be drawn about these features of DGBLL as well as the evaluation framework's usage and limitations. Both of these outcomes will be discussed in the subsequent discussion, leading ultimately to the identification of areas for further research.

4. Results

Based on the evaluation tool that was developed and implemented as described above, the analysis of 50 language learning applications was able to highlight common patterns and structures. While this examination is not comprehensive in terms of available applications, the trends identified in the aforementioned areas are indicative of the types of gamified educational programs currently available for language learning purposes. Given the relatively small sample size, numerical and correlational data are tentative, requiring further exploration, and are thus not reported here. The results do not encompass all of the collected data, but rather, focus on those items that are most pertinent for understanding the potential and limitations of current DGBLL. In particular, the selected results included here focus on the degree to which theories of DGBL, CALL, or language learning acquisition inform these applications.

4.1 Didactic analysis

The data suggests that the majority of DGBLL utilize behaviorist techniques to facilitate the acquisition and reinforcement of receptive competencies for learners who are beginning learners. The determination as to whether an application incorporates largely behaviorist or constructivist methods is not made on the basis of any one element, but rather, reflects an analysis of a variety of features regarding content, feedback, collaborative and cooperative opportunities, use of multimedia and authentic materials, and scaffolding.

Notable is the emphasis on receptive, over productive, skills, as indicated in Figure 1.

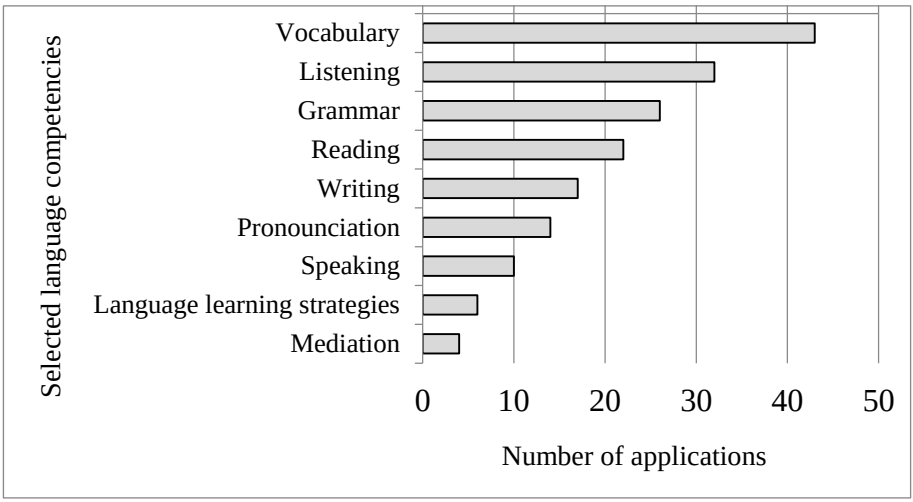

Figure 1: Language competencies

Of the 50 programs, 86% (43) incorporate or focus exclusively on the acquisition or recall of lexical items. These are most frequently proffered explicitly, as discrete terms with limited contextual embedding. While 64% (32) of the programs incorporate listening tasks, none enable learners to submit a free response in response to an audio prompt. Instead, comprehension of audio input is evaluated through closed items. Likewise, there are no applications that incorporate both speaking activities and constructivist elements. In the majority of applications, exercise-like activities predominate, as indicated in Figure 2.

The majority (86%) of applications utilize closed formats, although some construct these in ways that mimic open-ended activities by requiring learners to carry out a command or complete a task. In the application *Islands,* learners practice prepositions of place, moving a character *in front of, behind,* or *next to* a given object. Such tasks integrate the primarily behaviorist learning task into the game world. Other programs, however, do not achieve this interplay. *English Attack* is an example of such a program. By typing in the correct forms of irregular verbs, players are able to have their character, a comic-type animal, win a race. The connection between the content and the game is not evident.

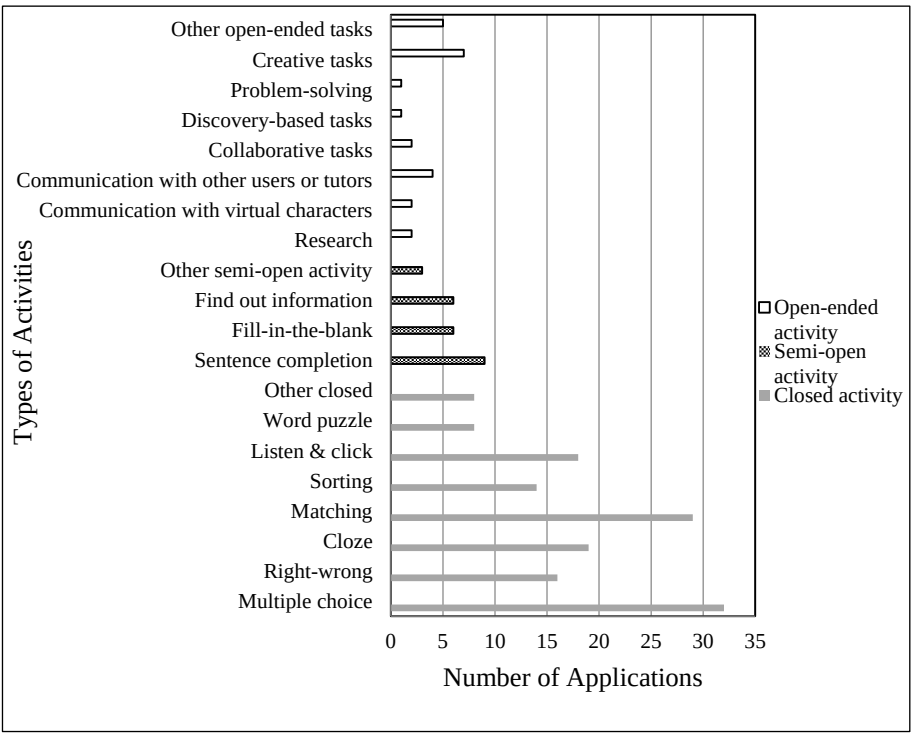

Figure 2: Activity types

Although half of the applications allow the user to choose the order in which to complete the given exercises, other elements that consider the program's adaptivity demonstrate that the majority of the language learning programs favor behaviorist instructional methods.

Table 4: Behaviorist elements

Element	Drill & practice	Pre-determined order	Pre-determined content	Immediate feedback	Repetition until correct
Percent	86%	50%	82%	86%	84%
Number of appl.	43	25	41	43	42

These elements are strongly informed by game mechanics. In all of the aforementioned cases, learners receive points, move up on a leaderboard, collect a bonus, or receive praise for correct responses, while incorrect responses are penalized accordingly.

The preeminence of behaviorist elements, while potentially a result of game design issues, is likely further influenced by the absence of foreign language learning expertise in the design and evaluation of the given programs. Given how much theoretical and empirical data exists regarding language learning and acquisition, an assessment of the degree to which this existing research informs available applications provides some indication of their alignment with these research-based principles.

The majority of the reviewed applications do not indicate any formal or informal associations with researchers or institutions of language learning. 42% (21) of the applications were developed with apparent input from experts in the fields of second language education, emerging from a cooperation with an academic institution, such as the *Goethe-Institut* (6%), as products of textbook companies (8%), or in collaboration with an academic in linguistics or a related field as a primary author (6%). This is not to say that the remaining 58% (29) do not reflect the expertise of academics in language acquisition, but their input in these cases is not apparent or cannot be ascertained. While the presence or absence of academic input or reviews does not, in itself, determine the nature of these products, it does suggest that knowledge of language pedagogy may be constrained. This assessment is reinforced by the apparent finite language knowledge of the producers, with 32% (16) of the applications revealing linguistic errors.

4.2 Didactic interactivity

While an analysis of the underlying didactic method overlaps to some degree with the concept of didactic interactivity, the focus in the former case is on using the presence or absence of interactive elements to form a description of the underlying learning theory. The emphasis in examining the latter case is on the ways in which the application facilitates learner autonomy (cf. Jones, Stuhlmann & Zeyer 2016).

Only 12% (6) of the applications collect data or conduct a pre-test to assess users' pre-existing language skills and adapt instruction targeted to their level of ability. Of these 6 applications, half (3) automatically select a level for learners to begin within the program. For the remainder, the learner is directed to make a selection regarding the subsequent level of difficulty based on these results, but is free to choose otherwise. None of the applications automatically change the level of difficulty based on the user's responses; repeated playing results in the same set of items. While the sequence of items might vary through repeated attempts, there is no indication that this correlates to an analysis of the item's difficulty for the individual user or to a standardized assessment of the item's level of challenge (i.e. via word lists or order of introduction for grammar items).

In addition to selecting the level of difficulty, users can select, in 30% (15) of the applications, from several available topics or create their own input. For those applications in which the content is predetermined, the topics are arranged according to the level of difficulty, so that a beginning learner, for example, who chooses to focus on content related to "dining out" will necessarily have to select more challenging input, at least based on the application's assumption of what constitutes more sophisticated language. Despite branding the levels in terms of increasing difficulty, in many cases, what changes are simply the vocabulary words; other linguistic competencies and activities and tasks do not noticeably change to reflect greater facility with either the language or the gameplay. None of the applications have the ability to accommodate individual preferences regarding learning styles, prior gameplay experience, or personal interests.

Feedback in the majority of applications is focused on closed or semi-closed items. While users can select, 16% (8) of the time, whether they want acoustic signals (available in a total of 50% of the programs) in addition to visual reactions (80% of the programs), they cannot otherwise adapt the type, depth, or timing of the feedback. In 10% (5) of the applications, learners can click on the feedback to indicate they want more information. However, in none of the cases does the elaborated feedback address learner errors beyond a standard correction (i.e. "Here is how to spell brother ") or statement of a rule (i.e. "The present participle is created with a form of *to be* + *-ing* ").

The few programs (8%; 4) that provide feedback to open-ended items do so either through peer-learning structures or with the help of human tutors; in these cases, input and feedback are asynchronous. Despite the fact that one program advertises large user communities who can respond to requests for feedback virtually instantaneously, requests for feedback from multiple peers were ultimately left unanswered in 80% of the cases.

4.3 Game-based characteristics

The majority of the tested applications utilize a narrow range of common game mechanics and attributes. These most frequently found elements tend to emphasize discrete units of achievement with limited interactional quality.

The most commonly utilized mechanic is the use of points that are accumulated over the course of play. Also frequently found are progression indicators, status rankings, or opportunities to "level up". One or more of these mechanics could be found in 90% of the programs examined, as indicated in Figure 3.

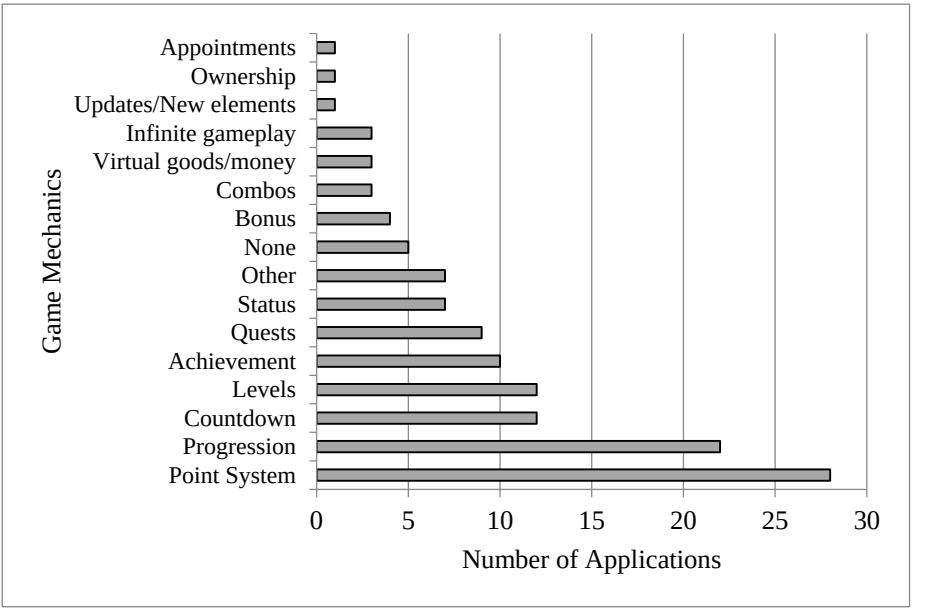

Figure 3: Game mechanics

These game-based rewards are used to both frame feedback and, theoretically, enhance learning, despite the questionable efficacy of doing so (cf. Abramovich, Schunn & Higashi 2013; Hughes & Lacy 2016)

Applications that utilize implicit rewards as a means of giving feedback are much less common, present in only 10% (5) of the applications. In these cases, the learner is not given an indicator of "right" or "wrong," but rather, consequences that indicate the (in)accuracy of the given response. In *Daumerlings Wanderschaft* (*Tom Thumb's Journey*), for example, the character's rapid demise suggests a false choice has been made.

5. Discussion

The proliferation of DGBLL would suggest that there is a wide variety of available applications that meet the varying needs of different types of learners. While the data indicates that there is, indeed, substantial variety among applications as regards, for example, narrative development, other indicators of variety and thus, appropriateness for unique learners, are largely absent.

The majority of DGBLL in the sample share common features that belie the impression of variety created by the sheer number of available applications. The

vast majority consist of simplistic content, behaviorist methods, and straightforward game mechanics, which reinforce one another to construct programs that largely target the receptive, lexical proficiencies of beginning language learners. Combined with straightforward rewards systems and intuitive gameplay, these programs are promising as tools for engaging students in opportunities to practice. However, programs that combine complex linguistic content and simulation-like play, narrative or inquiry that stimulate exploration or creative application are found infrequently (cf. Göbel, Wendel, Ritter & Steinmetz 2010).

Although the accumulating evidence suggests that those applications that create flow are most effective in terms of both affective and cognitive gains (Kiili, Freitas, Arnab & Lainema 2012), it apparently remains a design challenge to achieve this interplay. Many of the DGBLL address the challenge of incorporating academic content with gamification by, simply, not connecting them. Instead, the gamified aspects of the program serve as a reward for skills achieved, or as an incentive to practice a skill, regardless of the fact that such mechanics may be counterproductive in certain contexts (Landers 2014: 753).

6. Conclusion

The goal of this study was twofold. Initially, the aim was to develop a tool that could be utilized to evaluate dedicated DGBLL applications. Subsequently, the tool was applied to evaluate existing DGBLL. Conclusions from both components of this undertaking highlight the challenges with evaluating and designing DGBLL.

Despite a relatively narrow focus on games and gamified learning programs designed specifically for CALL, innumerable variations in design and pedagogy, and the interaction between the two, highlight the complexity of analyzing these products. While this evaluation tool offers a framework for considerations of DGBLL applications, more work needs to be done in order to guide potential users more precisely in their selection of appropriate tools. The tool itself likewise requires further testing and implementation, both to more fully assess its validity and to test its conclusions with a larger sample. Narrowing the focus, on the one hand, to certain types of applications, will validate trends already described here. Expanding the focus, on the other hand, and utilizing the tool with, for example, commercial products (COTS) will facilitate comparative analyses.

Despite the development of CALL theory from behavioristic to more integrative approaches, the majority of available games reflect the same activities and formats found in early CALL. Roche's (2003) analysis that technological advancement has led to pedagogical regression continues to be borne out. These

dressed-up drills have many advantages for autonomous learning. What they cannot do is simulate the dense intertwining of linguistic and pedagogical knowledge of effective language educators in interactive, communicative language classrooms.

Ultimately, applications that emphasize authentic skills over isolated ones and that embed tasks in sophisticated game structures and enable collaborative gameplay will be better able to engage both serious learners and enthusiastic gamers. Going forward, the goal for language learning games is to find the skilled matchmakers who can marry these complex language tasks with such sophisticated game mechanics.

Eingang des revidierten Manuskripts 13.11.2017

Applications mentioned

Babbel (2007). Lesson Nine GmbH.
Busuu (2008). Busuu Online S.L.
Daumerlings Wanderschaft (2013) Goethe-Institut.
Duolingo (2011). Duolingo.
English Attack (2012) Entertainment Learning.
Islands (2012) Pearson Education Limited.
Mingoville (2007). Savivo.

References

Aarseth, Espen (2014): I fought the law: transgressive play and the implied player. In: Segal, Naomi & Koleva, Daniela (eds.): *From Literature to Cultural Literacy:* Palgrave Macmillan UK, 180-188.

Abramovich, Samuel; Schunn, Christian & Higashi, Ross M. (2013): Are badges useful in education? It depends upon the type of badge and expertise of learner. *Educational Technology Research Development* 61: 2, 217-232 [DOI 10.1007/s11423-013-9289-2].

Anderson, Lorin W.; Krathwohl, David R.; Airasian, Peter W.; Cruikshank, Kathleen A.; Mayer, Richard E.; Pintrich, Paul R.; Raths, James & Wittrock, Merlin C. (2001): *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Pearson.

Anyaegbu, Ruphina; Ting, Wei & Li, Yi (2012): Serious game motivation in an EFL classroom in Chinese primary school. *TOJET: The Turkish Online Journal of Educational Technology* 11: 1.

Arnab, Sylvester; Lim, Theodore; Carvalho, Maira B.; Bellotti, Francesco; Freitas, Sara de; Louchart, Sandy; Suttie, Neil; Berta, Riccardo & Gloria, Alessandro de (2015): Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology* 46: 2, 391-411.

Bopp, Matthias (2006): Didactic analysis of digital games and game-based learning. In: Pivec, Maja (ed.): *Affective and emotional aspects of human-computer interaction* (2. ed.). Amsterdam: IOS Press, 8-37.

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

Burston, Jack (2003): Software selection: a primer on sources and evaluation. *CALICO Journal* 21: 1, 29-40.

Carvalho, Maira B.; Bellotti, Francesco; Berta, Riccardo; Gloria, Alessandro de; Sedano, Carolina I.; Hauge, Jannicke B.; Hu, Jun & Rauterberg, Matthias (2015): An activity theory-based model for serious games analysis and conceptual design. *Computer Education* 87, 166-181.

Chalak, Azizeh & Ahmadi, Behzad (2017): Integration of serious games in teaching English as a foreign language to Iranian children. *International Journal of Foreign Language Teaching & Research* 5: 17, 77-87.

Chapelle, Carol A. (2001): *Computer applications in second language acquisition: Foundations for teaching, testing and research.* Cambridge, UK: Cambridge University Press.

Chiu, Yi-hui; Kao, Chian-wen & Reynolds, Barry 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 [DOI 10.1111/j.1467-8535.2012.01295.x].

Conati, Cristina & Manske, Micheline (2009): Evaluating adaptive feedback in an educational computer game. In: Ruttkay, Zsófia; Kipp, Michael; Nijholt, Anton & Vilhjálmsson, Hannes H. (eds.): *Proceedings of IVA 2009, 9th International conference on intelligent virtual agents.* Switzerland: Springer, 146-158.

Connolly, Thomas; Stansfield, Mark & Hainey, Thomas (2009): Towards the development of a games-based learning evaluation framework. In: Connolly, Thomas M.; Stansfield, Mark & Boyle, Elizabeth A. (eds.): *Games-based learning advancements for multi-sensory human computer interfaces.* UK: IGI Global, 251-273.

Consalvo, Mia (2007): *Cheating: Gaining advantage in videogames.* Cambridge, Massachusetts: The MIT Press.

Consalvo, Mia (2011): Using your friends: social mechanics in social games. In: *Proceedings of the 6th International conference on the foundations of digital games.* New York: Association for Computing Machinery, 188-195.

Cornillie, Frederik; Thorne, Steven L. & Desmet, Piet (2012): Digital games for language learning: challenges and opportunities. *ReCALL* 24: 3, 243-256.

Crookall, David (2010): Serious games, debriefing, and simulation/gaming as a discipline. *Simulation & Gaming* 41: 6.

Deterding, Sebastian; Dixon, Dan; Khaled, Rilla & Nacke, Lennart (2011): From game design elements to gamefulness: Defining "Gamification". In: *Proceedings of the 15th International academic MindTrek conference envisioning future media environments.* New York, NY: ACM, 9-15.

Ellis, Rod (2005): Principles of instructed language learning. *System* 33: 2, 209-224 [DOI 10.1016/j.system.2004.12.006].

Feist, Silke (2008): Erwerb einer studienspezifischen Schreibkompetenz in der Fremdsprache am Beispiel des Sprachkurses Uni-Deutsch der Deutsch-Uni Online (DUO). *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 13: 1, 1-25.

Freitas, Sara de & Oliver, Martin (2006): How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers & Education* 46: 3, 249-264.

Gee, James P. & Hayes, Elisabeth (2012): Nurturing affinity spaces and game-based learning. In: Steinkuehler, Constance; Squire, Kurt & Barab, Sasha (eds.): *Games, learning, and society: Learning and meaning in the digital age.* Cambridge: Cambridge University Press, 129-153.

Göbel, Stefan; Wendel, Viktor; Ritter, Christopher & Steinmetz, Ralf (2010): Personalized, adaptive digital educational games using narrative game-based learning objects. In: Zhang, Xiaopeng; Zhong, Shaochun; Pan, Zhigeng & Yun, Ruwei (eds.): *Entertainment for education. Digital techniques and systems: 5th International conference on E-learning and games, edutainment 2010, Changchun, China, August 16-18, 2010*. Springer, 438-445.

Henry, Alastair (2013): Digital games and ELT: Bridging the authenticity gap. In: Ushioda, Ema (ed.): *International perspectives on motivation: Language learning and professional challenges*. London: Palgrave Macmillan, 133-155.

Hochleitner, Christina; Hochleitner, Wolfgang; Graf, Corniela & Tscheligi, Manfred (2015): A heuristic framework for evaluating user experience in games. In: Bernhaupt, Regina (ed.): *Game user experience evaluation.* Switzerland: Springer, 187-212.

Hubbard, Philip (1987): Language teaching approaches, the evaluation of CALL software, and design implications. In: Smith, William F. (ed.): *Modern media in foreign language education: Theory and implementation.* Hastings-on-Hudson, NY: National Textbook Company, 227-254.

Hubbard, Philip (1991): Evaluating computer games for language learning. *Simulation & Gaming* 22: 2, 220-223.

Hubbard, Philip (2006): Evaluating CALL software. In: Ducate, Lara & Arnold, Nike (eds.): *Calling on CALL: From theory and research to new directions in foreign language teaching.* UK: Computer Assisted Language Instruction Consortium, 1-26.

Hughes, Michael & Lacy, Jeff C. (2016): "The sugar'd game before thee": Gamification revisited. *portal: Libraries and the Academy* 16: 2, 311-326.

Hwang, Gwo-Jen & Wu, Po-Han (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. [DOI 10.1111/j.1467-8535.2011.01242.x].

Issing, Ludwig J. & Klimsa, Paul (eds.) (2002): *Information und Lernen mit Multimedia und Internet* (3. Aufl.). Weinheim: Beltz.

Jones, Roger D.; Stuhlmann, Sebastian & Zeyer, Tamara (2016): Interaktives Fremdsprachenlernen: Potenziale und Herausforderungen. In: Zeyer, Tamara; Stuhlmann, Sebastian & Jones, Roger D. (eds.): *Interaktivität beim Fremdsprachenlehren und -lernen mit digitalen Medien: Hit oder hype?* Tübingen: Narr Francke Attempto, 11-42.

Karppi, Tero & Sotamaa, Olli (2012): Rethinking playing research: DJ HERO and methodological observations in the mix. *Simulation & Gaming* 43: 3, 413-429.

Kerres, Michael & Bormann, Mark (2009): Explizites Lernen in Serious Games: Zur Einbettung von Lernaufgaben in digitalen Spielwelten. *Zeitschrift für E-Learning* 4.

Kiili, Kristian; Freitas, Sara de; Arnab, Sylvester & Lainema, Timo (2012): The design principles for flow experience in educational games. *Procedia Computer Science* 15, 78-91.

Kober, Christina; Paland-Riedmüller, Ines & Hafner, Stephanie (2015): "Daumen hoch" für das virtuelle Klassenzimmer. Zur Förderung mündlicher Interaktion in studienvorbereitenden Online-Sprachkursen durch den Einsatz eines virtuellen Klassenzimmers mit ergonomischer Benutzeroberfläche. In: Nistor, Nicolae & Schirlitz, Sabine (Hrsg.): *Digitale Medien und Interdisziplinarität.* Münster: Waxmann, 228-238.

Landers, Richard N. (2014): Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & Gaming* 45: 6, 752-768.

Leakey, Jonathan (2011): *Evaluating computer-assisted language learning: An integrated approach to effectiveness research in CALL*. Oxford: Peter Lang.

Leutner, Detlev (2002): Adaptivität und Adaptierbarkeit multimedialer Lehr- und Informationssysteme. In: Issing, Ludwig J. & Klimsa, Paul (eds.), 115-126.

Levy, Mike & Stockwell, Glenn (2006): *CALL dimensions: Options and issues in computer assisted language learning.* Mahwah, N.J.: Lawrence Erlbaum.

Lombardi, Ivan (2012): Computer games as a tool for language education. *G|A|M|E Games as Art, Media, Entertainment* 1: 1.

Mayer, Igor; Bekebrede, Geertje; Harteveld, Casper; Warmelink, Harald; Zhou, Qiqi; van Ruijven, Theo; Lo, Julia; Kortmann, Rens & Wenzler, Ivo (2014): The research and evaluation of serious games: Toward a comprehensive methodology. *British Journal of Educational Technology* 45: 3, 502-527.

McMurry, Benjamin L.; West, Richard E.; Rich, Peter J.; Williams, David D.; Anderson, Neil J. & Hartshorn, K. J. (2016): An evaluation framework for CALL. *TESL-EJ* 20: 2.

Moreno-Ger, Pablo; Burgos, Daniel; Martínez-Ortiz, Iván; Sierra, José L. & Fernández-Manjón, Baltasar (2008): Educational game design for online education. *Computers in Human Behavior* 24: 6, 2530-2540.

Peachey, Anna & Childs, Mark (2011): Virtual Worlds and Identity. In: Peachey, Anna & Childs, Mark (eds.): *Reinventing ourselves: Contemporary concepts of identity in virtual worlds.* London: Springer, 1-12.

Reeder, Kenneth; Heift, Trude; Roche, Jörg; Tabyanian, Shabaz; Schlickau, Stephan & Gölz, Peter (2004): Toward a theory of e/valuation for second language learning media. In: Fotos, Sandra & Browne, Charles (eds.): *New Perspectives on CALL*. Mahwah, NJ: Lawrence Erlbaum, 255-278.

Reinders, Hayo & Wattana, Sorada (2011): Learn English or die: The effects of digital games on interaction and willingness to communicate in a foreign language. *Digital Culture & Education* 3: 1, 3-29.

Reusser, Kurt (1996): From cognitive modeling to the design of pedagogical tools. In: Vosniadou, Stella; De Corte, Erik; Glaser, Robert & Mandl, Heinz (eds.): *International Perspectives on the design of technology-supported learning environments.* Mahwah, NJ: Lawrence Erlbaum, 81-103.

Ritterfeld, Ute & Weber, René (2006): Video games for entertainment and education. In: Vorderer, Peter & Bryant, Jennings (eds.): *Playing video games: Motives, responses, and consequences.* New Jersey: Lawrence Erlbaum Associates, 399-413.

Roche, Jörg (2003): Plädoyer für ein theoriebasiertes Verfahren von Software-Design und Software-Evaluation. *Deutsch als Fremdsprache* 40: 2, 94-103.

Salen, Katie & Zimmerman, Eric (2003): *Rules of play: Game design fundamentals.* Cambridge, Mass.: MIT Press.

Saltzman, Marc (2000): *Game Design. Die Geheimnisse der Profis (X-Games)*. Burgthann: Markt+Technik.

Schmidt, Torben; Schmidt, Inke & Schmidt, Philipp R. (2016): Digitales Spielen und Lernen – A perfect match? Pädagogische Betrachtungen vom kindlichen Spiel zum digitalen Lernspiel. In: Dadaczynski, Kevin; Schiemann, Stephan & Paulus, Peter (Hrsg.): *Gesundheit spielend fördern: Potenziale und Herausforderungen von digitalen Spieleanwendungen für die Gesundheitsförderung und Prävention*. Weinheim: Beltz Juventa, 1-33.

Shute, Valeria (2008): Focus on formative feedback. *Review of Educational Research* 78: 1, 153-189.

Sicart, Miguel (2008): Defining game mechanics. *Game Studies – the International Journal of Computer Game Research* 8:2 [Online: http://gamestudies.org/0802/articles/sicart 14.10.2017].

Strzebkowski, Robert & Kleeberg, Nicole (2002): Interaktivität und Präsentation als Komponenten multimedialer Lernanwendungen. In: Issing, Ludwig J. & Klimsa, Paul (eds.), 229-246.

Sykes, Julie M. (2013): Synthetic immersive environments and second language pragmatic development. In: Chapelle, Carol A. (ed.): *The encyclopedia of applied linguistics.* Malden, Massachusetts: Wiley-Blackwell.

Sykes, Julie M. & Reinhardt, Jonathon (2013): *Language at play: Digital games in second and foreign language teaching and learning.* Boston, Mass.: Pearson.

Teddlie, Charles & Yu, Fen (2007): Mixed methods sampling; a typology with examples. *Journal of Mixed Methods Research* 1: 1, 78-100.

Tomlinson, Brian (2010): Principles and procedures for self-access materials. *Studies in Self-Access Learning Journal* 1: 2, 72-86.

Tomlinson, Brian (2014): Materials evaluation. In: Tomlinson, Brian (ed.): *Developing materials for language teaching.* London, England: Bloomsbury, 15-36.

Vesselinov, Roumen & Grego, John (2012): *Duolingo effectiveness study* [Online: http://static.duolingo.com/s3/DuolingoReport_Final.pdf 14.10.2017].

Vesselinov, Roumen & Grego, John (2016a): *The Babbel efficacy study final report* [Online: https://press.babbel.com/en/releases/downloads/Babbel-Efficacy-Study.pdf 14.10.2017].

Vesselinov, Roumen & Grego, John (2016b): *The Busuu efficacy study final report.* New York. [Online: http://comparelanguageapps.com/documentation/The_busuu_Study2016.pdf 14.10.2017].