Language Testing in Asia

REVIEW Open Access

A case for the use of the ability-in language user-in context orientation in game-based assessment

Alexandra Lay*, Elizabeth Patton and Micheline Chalhoub-Deville

* Correspondence:
almart25@uncg.edu
Department of Educational Research Methodology, University of North Carolina Greensboro, Greensboro, NC 27406-6170, USA

Abstract

Dynamic assessments in general, and game-based assessment (GBA) specifically, compel us to rethink prevailing language testing conceptualizations of context. Context has traditionally been portrayed with a cognitive orientation, which focuses on static abilities, ignores complex interactions, devalues the role of tasks in determining scores, and makes connections to learning potential difficult. *Ability-in language user-in context* lays the groundwork for a new framework for use in language testing and GBA by shifting the conversation of interaction and context to a level of entanglement not yet considered in the field. The paper makes connections to concepts in other disciplines (e.g., intensive and extensive context, reciprocal interaction, and co-construction) and problematizes relevant design, measurement, and validity considerations. We identify claims in the areas of generalizability, score comparability, and dimensionality and suggest areas for future research. We end the article by inviting language testers in general, and those in the Asian language testing community in particular, to engage in GBAs.

Keywords: Ability-in language user-in context, Reciprocal interaction, Context, Game-based assessment, Dynamic assessment

Background

The field of second language acquisition is developing interest in the utilization of games as second language (L2) learning tools (e.g., Rankin et al. 2006; Rankin et al. 2008; Ranalli, 2008). The use of L2 testing games, however, is nascent. It is critical, at this early stage of research and development, for those interested in game-based assessments (GBA) to consider some fundamental interaction and contextual concepts that drive how we pursue our assessments. The type of interaction and role of context have long been debated in the field of language assessment (Bachman 1990; Bachman and Palmer 1996; Chalhoub-Deville 2003; Chalhoub-Deville and Deville 2006; Chalhoub-Deville 2009, Deville and Chalhoub-Deville 2006; Purpura 2008). These same issues are beginning to spark discussion within the general area of GBAs (DiCerbo 2014; DiCerbo et al. 2016; Mislevy et al. 2012; Mislevy et al. 2015), which we position within a broader category of dynamic assessment. Traditionally, context and interaction have been portrayed with a cognitive orientation. We contend, however, that *ability-in language user-in context*, introduced by Chalhoub-Deville (2003), offers a more apt orientation. Ability-in language user-in context lays the groundwork for a new framework for use in language testing and GBAs by shifting the conversation of interaction and context to a level of entanglement not yet considered in the field.

The present paper explores the role of context and interaction in dynamic assessments, specifically GBA, and the utility of ability-in language user-in context as a guiding perspective. The paper makes connections to key concepts that appear in other disciplines (e.g., intensive and extensive context, reciprocal interaction, and co-construction) and problematizes relevant design, measurement, and validity considerations. We identify claims in the areas of generalizability, score comparability, and dimensionality and suggest areas for future research. The paper explores the following broad research questions:

1) How does ability-in language user-in context depict concepts such as context and interaction?
2) How do the concepts of context and interaction figure in prominent language testing and educational measurement models?
3) How does ability-in language user-in context relate to dynamic assessments such as GBAs?
4) What are some relevant measurement considerations when considering GBAs through an ability-in language user-in context orientation?

Dynamic assessment and language testing: context and interaction

Dynamic assessments, a particularly complex class of interactive assessments, have recently received increasing attention in response to rising criticisms of conventional, static assessments, such as their inadequacy in assessing children's cognitive capacities (Tzuriel 2001). While conventional assessments typically emphasize standardized contexts and fixed interaction, dynamic assessments underscore interactive contexts, which facilitate a fluid, complex transaction between *person* and *task*. Dynamic assessment places a strong emphasis on the *processes* used throughout assessment as well as the *modifiability* of those processes (Haywood et al. 1992). Dynamic assessment underscores a documentation of students' potential for learning. The conceptual basis for dynamic assessment may be attributed to the lifelong work of developmental and cognitive psychologists such as Vygotsky and Feuerstein (Tzuriel 2001). Digital games can offer a rich example of dynamic assessments given the connections they create between learning and assessment. Digital games scaffold players' interaction and provide them with relevant information to promote game progression. Digital games, and specifically GBAs, offer complex interactions of person/player and task. These complex interactions are at the heart of what we deliberate when we discuss issues of context.

Discussions of context, i.e., person and task interaction, in language testing were made prominent by Bachman (1990) and Bachman and Palmer (1996). "Bachman (1990) composed a general interactionalist L2 [second/foreign language] construct definition, which includes 'both knowledge, or competence, and the capacity for implementing, or executing that competence in language use' in context (Bachman 1990: 84)" (Chapelle 1998, p. 44). In writing about Communicative Language Ability (CLA), Bachman and later Bachman and Palmer describe attributes that test takers bring to an assessment situation such as their topical knowledge, language knowledge, and personal characteristics. They also outline task features relevant to a given testing situation by using the target language use domain (TLU) checklist. Similarly, Chapelle (1998) advocates for an interactionalist approach to assessing language ability. She argues for "an examination of the construct …where the relevant attributes of context, credited for influencing variation in response patterns, are carefully specified as part of the construct" (as cited in Purpura 2008, p. 6).

Chalhoub-Deville (2009) argues that the interactional approach advanced by researchers such as Chapelle, Bachman, and Palmer is "predominantly cognitive," depicts communication "largely as static and fixed," and "emphasizes the ability within an individual because ultimately we are interested in awarding individual scores" (p. 253). Chalhoub-Deville (2009) and Schwabe et al. (2016), using Snow's (1994) definition, characterize the type of interaction discussed by Chapelle (1998 and as cited in Purpura, 2008), Bachman (1990), and Bachman and Palmer (1996) as "interdependent interaction," where person and task variables are separate but related entities. Deville and Chalhoub-Deville (2006) argue that a "focus on a stable core of abilities, which has been the usual practice in testing, affords necessary but not sufficient information about the interactional nature of […] contexts, and ultimately about score interpretability" (Chalhoub-Deville 2009, p. 257). The manner in which researchers such as Bachman (1990), Bachman and Palmer (1996), and Chapelle (1998 and as cited in Purpura, 2008) describe person and task interaction ultimately falls short in terms of describing what occurs in a complex interactional assessment situation, such as what may be observed in dynamic GBAs. Chalhoub-Deville (2009) and Schwabe et al. (2016) favor "reciprocal interaction," which regards person and task as inseparable entities that change one another over time.

Researchers in dynamic assessment advance notions similar to "reciprocal interaction" albeit under a different label, i.e., "transactional" (Haywood et al. 1992). Haywood et al. (1992) characterize the transactional perspective "by the *reciprocal* effects of all components (factors) and by the complex circular process" (p. 50, italics added). They add that this "implies a dynamic relationship among subject, assessor, materials, and tasks, such that each influences the others" (p. 52). Despite the varying and confusing terminology (reciprocal interaction vs. transactional perspective), the inherent concepts are commensurate. Both reciprocal interaction and the transactional perspective center on the idea that major variables in an assessment (person, task, etc.) are in constant interaction, working in tandem to change one another.

Ability-in language user-in context: connections to concepts in related disciplines

With the introduction of ability-in language user-in context, Chalhoub-Deville (2003) begins the shift from the more conventional consideration of interaction (interdependent interaction) to a more progressive, dynamic consideration (reciprocal interaction), and in turn, places greater emphasis on the role of context in regard to both task and person. Ability-in language user-in context revolves around the idea that in a given language assessment situation, the construct of second language ability is in reciprocal interaction with assessment tasks, which produces a specific performance on that assessment. This conceptualization follows arguments advanced in applied linguistics such as Kramsch's (1986, 1998), notion of co-construction:

> Whether it is face-to-face interaction between two or several speakers or the interaction between the reader and the written text, successful interaction presupposes not only a shared knowledge of the world, the reference to a common external context of communication, but also the construction of a shared internal context or "sphere of inter-subjectivity" that is built through the collaborative effort of the interactional partners (1986, p. 367).

Kramsch contends that successful interaction necessitates a reference to a shared external context and a co-construction of a joint internal context. In such a conceptualization of context, language interaction is entangled, i.e., in assessment terms, person, and task are to be regarded as inextricable. As with arguments in dynamic assessment, we cannot divorce a student's ability from the larger context in which the knowledge was acquired.

In concert with Kramsch's notions of context, and as Fig. 1 illustrates, Chalhoub-Deville (2009) depicts an association between intensive and extensive contexts and suggests locating the *Assessment Situation* at their intersection. Figure 1 situates a person's/student's ability within societal norms and cultural beliefs, i.e., the extensive context. By adding elements of culture, schooling, and society to the educational assessment context, we are drawing on discussions such as Gilbert's (1992) "extensive context" (in Chalhoub-Deville 2009) and Bronfenbrenner's (1979) "macrosystem" from his ecology of human development. The intensive context, which is also situated within the extensive context during assessment, describes the proximate sphere of the reciprocal person and task interaction. The intensive aspect of context speaks to the rich task-based interactions, which includes interactants, communication goals, resources available, cultural notions held, etc. The intersection of extensive and intensive contexts portrays how persons' communal resources such as values, schooling, and norms come to bear on the immediate scope of interaction with tasks. This is what we envision to be the case in an assessment situation.

Figure 2 is an outcome of the debate over the centrality of construct versus content, as elaborated in Chalhoub-Deville (2009). At the heart of the argument, we propose, is a continued commitment to constructs but, as the figure shows, the construct is enmeshed with the task and its content attributes. This entanglement of the ability, which traditionally is said to reside within a language user, and the task features is summarized in the expression: an ability-in language user-in context orientation. The Assessment Situation embodies the person or language use and the assessment tasks as

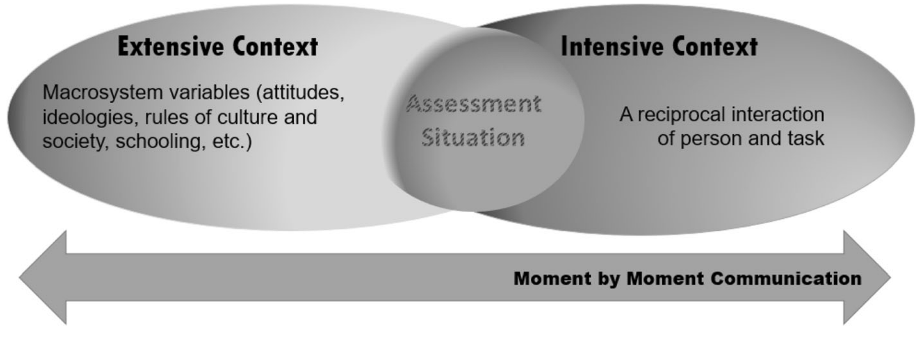
Extensive Context
Macrosystem variables (attitudes, ideologies, rules of culture and society, schooling, etc.)
Assessment Situation
Intensive Context
A reciprocal interaction of person and task
Moment by Moment Communication

Fig. 1 Representations of context (an adaptation from Chalhoub-Deville 2009, p. 253)

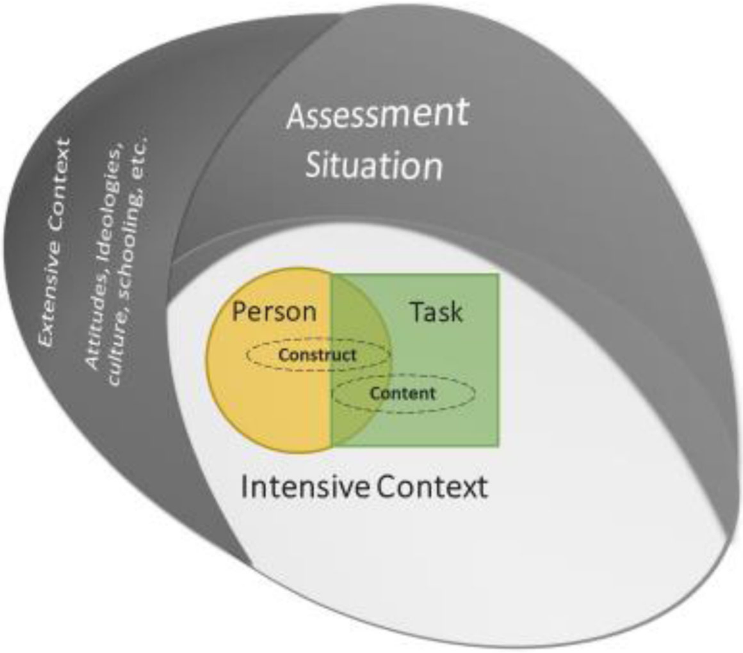

Fig. 2 A representation of context and person × task interaction as depicted in ability-in language user-in context

they exist within their given cultural structures, group attitudes and ideologies, schooling experiences, family engagement, community resources, etc.

In this Fig. 2 depiction, a person's participation in the assessment situation should not be a matter of mechanical communication and language regurgitation. Language ability/use is dynamic and structured within a person's extensive context elements. Moreover, a person embodies their own set of personal characteristics, knowledge of the construct, command of content areas, engagement in topics, among others that shape assessment performance. Parallels can be drawn here to the test taker attributes included in CLA (Bachman 1990; Bachman and Palmer 1996). However, characteristics and knowledge represented in a person's performance are not necessarily fixed or generic. This is evident in the literature we discuss throughout this paper, which richly documents variable interactional performance and co-construction and supports a more nuanced, interactive, and fluid portrayal of a person's ability.

Tasks may be said to represent an activity or exercise such as in performance or game-based assessments. Tasks have long been discussed in the second language acquisition (SLA) literature as critical activities that promote interlanguage development in environments that simulate real-world learning (e.g., Krahnke 1987; Long and Crookes 1992). In language testing, Bachman (1990) and Bachman and Palmer (1996) present a framework, known as the target language use domain (TLU) checklist, to document task features. The TLU checklist has roots in the test method literature. Typically, tasks, in the tradition of test method, are viewed as "necessary evil" tools needed to access a person's construct. In such a tradition, a task's influence needs to be minimized since we cannot eliminate it. In the present article, we take a different view of tasks and construct. We consider tasks, in the tradition of dynamic assessment, to represent real-world learning and to be integral to how the construct is to be portrayed. Task features are largely a function of the interaction between the person and task within the assessment situation. Therefore, our conceptualization of L2 ability is an entanglement of person and task.

Ability-in language user-in context: design and claims

We have presented ability-in language user-in context as an orientation in assessment that depicts a reciprocal person and task interaction. To further exemplify how this orientation differs from traditional measurement frameworks, i.e., interdependent person and task interaction, we will now turn our attention to the work of Mislevy. We have chosen to center our discussion on Mislevy and evidence centered design (ECD) (Mislevy et al. 2003) due to the prominence of ECD in the measurement literature. Mislevy's thinking has shaped influential educational achievement testing systems such as Race to the Top Consortia testing Program (Flowers et al. 2015), language testing programs, e.g., Next Generation TOEFL—now called iBT TOEFL (Chapelle et al. 2008), and innovative research projects (e.g., GBA). His contributions stand out because of their systematic attention to validity at the design and development level.

While ECD has several elaborate design layers, from Domain Analysis to Assessment Delivery, the most relevant to this discussion are the entities of the Conceptual Assessment Framework (CAF). Mislevy et al. (2015) states:

> The Conceptual Assessment Framework houses the Student Model, which describes aspects of the student (knowledge or skills) that need to be assessed, the Task Model, which describes the features of the assessment tasks, and the Evidence Model, which is the "bridge between what we see students do in situations (as described in task models) and what we want to infer about their capabilities (as expressed in [the student model])" (p. 30).

Interaction, as conceptualized in ECD, can be said to embrace an interdependent perspective. To explain, while ECD seeks to support through design, development, and analysis, claims about students' ability using tasks relevant to a situation of interest, it ultimately attempts to support inferences that move beyond the person in a given situation interacting with a specific task. This is more explicitly evident in the representation of claims as shown in Fig. 3.

As evidenced by Fig. 3, the task and student are ultimately separated resulting only in the "claim about student." ECD seeks to develop inferences regarding students'/persons' abilities, which are informed by, but not restricted or defined by, the task and the interaction. The inference of interest is ultimately the construct or underlying student abilities. This position is akin to Chapelle et al.'s (2008) claim representation with regard to the iBT TOEFL project (see Fig. 4). Similar to ECD, Chapelle et al. (2008), underscore the role of task and situations in engaging students and driving performance observations. Both, nevertheless, ultimately seek to render the complex relationships represented in these interactions to statements about abilities separate from the tasks or interactional situations. As Figs. 3 and 4 show, the grounds for the claims are based on a separation of the person (student, performance, construct, ability) from the task. Again, this representation follows the interdependent interaction mentioned earlier of conceptualizing tasks as the "necessary evil" method to get at needed

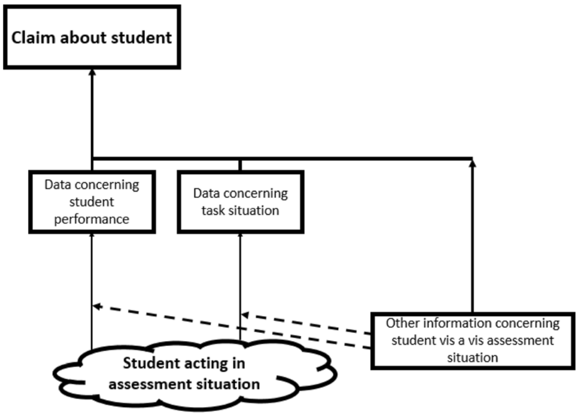
Claim about student
Data concerning student performance
Data concerning task situation
Other information concerning student vis a vis assessment situation
Student acting in assessment situation

Fig. 3 Selected representation of "an assessment design argument" (Mislevy et al. 2015, p. 28). Note: this version of the visual does not include the elements of the descriptive argument as found in the original, please refer to the article cited for the full visual

unobservable features. Figure 5 presents a reciprocal interaction, specifically ability-in language user-in context, which posits an integrated claim perspective. It emphasizes the inseparability of the task and the ability at both levels—the grounds and the inference about a student's ability. (Figs. 4 and 5, taken from Schwabe et al. (2016).) The distinction we highlight here impacts test design as well as score interpretation/inferences.

Established language testing and measurement models devote a great amount of attention to appropriately modeling student interaction with tasks in a given Assessment Situation. The task, however, is an auxiliary tool to externalize ability features of interest, obtain individual scores, afford generalizable ability interpretations, etc. As educational measurement and language assessment professionals embrace the use of more complex test tasks, incorporate technology to allow for collaborative test interactions, experiment with nontraditional test administration conditions, shift from documenting past achievement to characterizing students' potential for learning, which are all critical elements of dynamic assessment, consideration should be given to the increasingly complex and inextricable nature of person and task interaction. Next, we revisit dynamic assessment, focusing specifically on game-based learning and GBAs. We will articulate reasons why a reciprocal ability-in language user-in context orientation provides a more apt conceptual fit for the claims and inferences inherent in dynamic assessments. Finally, we will again consider that interaction at the reciprocal level

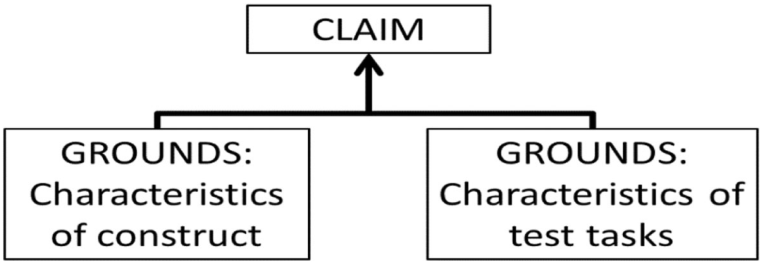
CLAIM
GROUNDS: Characteristics of construct
GROUNDS: Characteristics of test tasks

Fig. 4 An adaptation of the Chapelle et al. (2008) dual grounds for iBT TOEFL® inference from Schwabe et al. (2016)

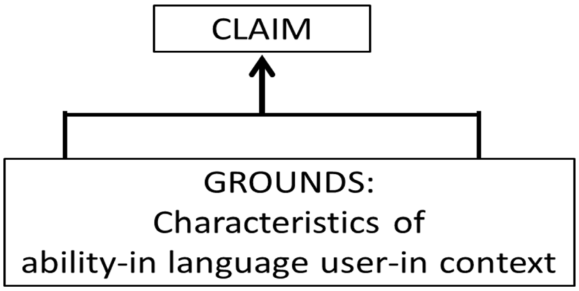

Fig. 5 An integrated inference perspective, framed as ability-in language user-in context from Schwabe et al. (2016)

refutes the claim that ability can be neatly separated from task in the assessment situation (i.e., scoring, interpretation, and use processes) and highlight the unique measurement challenges it presents.

Game-based learning: making the case for GBA

Digital games have the potential to enrich our repertoire of dynamic assessment. GBAs are frequently coupled with learning and teaching because data can be used to provide scaffolding to players, adjust their route through the game, and tailor the feedback they receive to allow progression. In this and the following section, we draw on the nascent research and development efforts in game-based learning and assessment to highlight the need for a reorientation to the nature of construct that dominates measurement and language assessment theory and practice. We start by examining the state of affairs in terms of L2 game-based learning. Positive documentation of L2 game-based efforts in the learning arena buttresses the case for engagement on the part of language testing and measurement professionals.

In the L2 field, some literature already exists that documents the contribution of digital games to learning (Hsu et al. 2015). Research on games as learning tools for English as a Foreign Language (EFL) students is quite promising. For example, Rankin et al. (2006) designed a study in which students at varying levels of “second language mastery” were asked to participate in game play intended to promote acquisition. The study concluded that those students at higher levels of mastery benefited from the game play much more so than those at lower levels of mastery. The authors suggest that students at the lower levels may have experienced “cognitive overload” trying to manage between the complex, dynamic aspects of the game with limited English language skills.

Additionally, Rankin et al. (2008) conducted a complex study where EFL students were randomly assigned to one of two groups: independent game play or collaborative play. Students in the collaborative game play completed tasks with group members, including two native English speakers. Results from an ANOVA indicated that those students who collaborated with native English speakers scored significantly higher on their post-tests than their independent game play counterparts ($p = .01$). Research seems to consistently suggest that the utilization of games can lead to increased second language vocabulary acquisition (Rankin et al. 2006; Rankin et al. 2008; Ranalli 2008).

Though research into EFL game-based learning in Asia is limited, the results are positive. Chiu et al. (2012) conducted a meta-analysis utilizing both fixed and random effects hierarchical linear models on 14 studies of EFL game-based learning in Taiwan, South Korea, and Hong Kong. General results indicated a medium positive effect size when comparing game-based learning to traditional instructional approaches (fixed effects model: Cohen's d = 0.674, CI = 0.550 to 0.797; random effects model: Cohen's d = 0.528, CI = 0.197 to 0.859). However, they found that that the degree to which games offer dynamic and interactive learning environments impacts their effectiveness as learning tools. Specifically, context-rich games had a large positive effect size (fixed effects model: Cohen*'s* d = 1.105, CI = 0.896 to 1.314; random effects model: Cohen's d = 0.844, CI = 0.006 to 1.683) versus drill and repeat games which yielded a small positive effect size (fixed effects model: Cohen*'s* d = 0.442, CI = 0.288 to 0.595; random effects model: Cohen's d = 0.406, CI = 0.162 to 0.650) on learning. Their results stress the importance of understanding the intricacies of a person by task interaction within complex and dynamic digital environments as these are more likely the types of games to receive research and funding in the future.

Game-based learning for EFL students has been found to be more effective in terms of closing the achievement gap between disadvantaged students and their peers (Hung et al. 2015). Hung and colleagues created and administered a cooperative crossword game called the Crossword Fan-Tan Game to 30 Taiwanese sixth grade students. Correspondingly, a wireless interface was designed which allowed teachers to monitor students' progress and display word maps for classroom discussion and peer analysis. Utilizing ANCOVA analysis, they found that low-achievement students utilizing the Crossword Fan-Tan Game had better learning outcomes than peers playing conventional games ($p = 0.01$). Additionally, interviews revealed that students utilizing the Crossword Fan-Tan Game reported increased focus ($p = 0.03$) and were more likely to look forward to playing the game again ($p = 0.044$) than peers playing conventional games.

The results reported in the nascent game-based L2 learning literature support positive learning potential and encourage us to take on explorations of GBAs. GBAs, however, present challenges that demand innovative engagement in terms of design and development as well as validation. These types of dynamic assessments contest traditional representations of person and task interaction and compel considerations of entangled constructs, much as depicted in reciprocal interaction. At the heart of the assessment and measurement considerations that we will take up later in the paper are issues such as scoring challenges, concerns about dimensionality, and generalizability limitations.

Game-based assessment: a reciprocal construct

Having provided promising research to support game-based learning, we turn our attention to assessments embedded within digital games. These are related to online learning tutors and technologically enhanced items but maintain unique features such as avatar creation, complex scenarios, and extended play. The goal of a digital game can include entertainment, learning and assessment. Examples include SimCityEDU (EA, Glasslab, Pearson), Simlandia (Nelson et al. 2011) and Newton's Playground (Ventura and Shute 2013). We limit our discussion to games, which incorporate complex scenarios rather than drill and repeat games such as Math Blaster's. Our ability to analyze complex data has increased with the application of data mining techniques to educational data, thus moving us past more simplistic games and GBAs. Additionally, results such as those reported by Chiu et al. (2012) highlight the increased efficacy of highly interactive dynamic environments for use in EFL contexts.

GBAs provide an exemplar of how technology is being used to extend thinking about what constitutes an assessment and their mode of delivery. GBAs have the potential to exemplify dynamic assessments that document interactants' L2 ability and their readiness to engage at increasingly complex levels. According to Popp et al. (2016), with the advent of animation and game-like interfaces, testing has gone from a tedious, simplistic, static, single-sensory experience to a dynamic, tailored, and multisensory experience. Within games, tasks are dynamic and change over time, thus altering players' experiences (Perez et al. 2016; Rueda et al. 2016). Trace data allows for tailored feedback to be given during game play, facilitates individualized, complex assessment experiences, and focuses on the next level of attainment (Koenig et al. 2010; Rupp et al. 2010).

Digital games provide opportunities to observe students in complex situations, which are closer to real world scenarios (DiCerbo 2014; Shute and Ke 2012). The desire for high fidelity/authentic assessment opportunities is not new to the measurement community, which increasingly has been experimenting with performance assessments. Some notable differences exist, however, between performance-based assessments and GBAs. For example, while tasks within a game may take a longer duration to complete there is no reason to restrict the number of tasks to the point that generalizability is threatened, a typical concern in performance assessments (Kane et al. 1999), as games can be played over many days and even months. Another distinctive feature of games is their ability to introduce *stealth assessment*, which according to Sireci and Faulkner-Bond, is "one of the most 'futuristic' phrases one can utter" (Sireci and Faulkner-Bond 2016, p.444). Stealth assessments refer to assessments which are "woven directly and invisibly into the fabric of the learning environment" (Shute and Ke 2012, p. 53).

While GBAs are a promising method of dynamic assessment, a review of the published literature shows that L2 GBA research and development does not match game-based L2 learning engagement. A survey of the published literature uncovers practically no publications focusing on L2 GBA. This is perhaps not surprising given the recent engagement in this domain, including in L2 learning—and testing typically follows versus leads in exploring and adopting innovations. Additionally, the measurement literature also identifies only a handful of publications (e.g., Mislevy et al. 2012; DiCerbo 2014; Mislevy et al. 2015; DiCerbo et al. 2016). Nonetheless, explorations of the scant GBA literature available in the wider measurement literature can be informative.

Educational Testing Service (ETS) houses some innovative research and development projects. One of these projects is Tetralogue, which focuses on data collected from 500 dyads involved in science assessment. The research investigates problem solving collaborations between two test takers while also interacting with two computer avatars (Hao et al. 2015). Psychometric and statistical analyses are being employed to analyze long time series of responses to quantify individual as well as dyad collaborative performances. Another ETS research project focuses on English language learners' using languages at their disposal, i.e., translanguaging, to demonstrate their content knowledge (Lopez et al. in press).

Outside of the language assessment literature, DiCerbo (2014) utilized the game Pop-Tropica to investigate persistence in performance. Persistence is measured in terms of the player being sufficiently motivated to complete a sufficiently difficult task in a chain of tasks that would lead to the completion of a quest. Thus, the appropriateness of the task as a measurement indicator is a function of the interaction between person and task. DiCerbo focuses only on the difficulty of a task as an attribute which can impact player persistence. However, other salient task features, for example, design of the game space, the game narrative, and the type of task (e.g., co-operational versus single player) could be considered.

Craig et al. (2015) utilize a game, *Zoo U*, for a cross cultural assessment of social skills in fourth and fifth graders. Previous literature informed their hypothesis that students from different countries, Japan versus the USA, would perform differentially on the assessment given differential cultural notions of what constitutes desirable social skills, e.g., subtle communication versus direct. Effectively, hypothesizing that extensive context factors would impact how students performed on social skills tasks. Their hypothesis was confirmed by empirical evidence. The researchers conclude with a recommendation to update scoring algorithms and definitions of construct when adapting GBAs for cross-cultural use.

The research projects and studies cited provide instructive examples of research and development of the next generation of technology-based dynamic assessments. These testing systems underscore, among other design and development elements, complex interactions, progressive levels of performance, single and cooperative responses, extended-time involvement, sustained engagement, multilingual and multicultural communication, as well as attention to process information. They are characterized by their highly interactive interfaces and the freedom they allow to players to choose their path through the game/assessment, thus moving away from the homogenization that has typified traditional assessments.

The features observed with L2 game-based learning and GBA are in sharp contrast to how traditional assessments are conceptualized, designed, and administered. Due to the complex and dynamic features inherent to GBA design (e.g., tasks, environments, and contexts), GBA environments are necessarily tied to contextual variables and make it challenging to disentangle ability features from the contexts/tasks. This entanglement is characteristic of a reciprocal interaction perspective, which is represented in ability-in language user-in context orientation. The validation process, specifically the nature of the claims, also needs to be adjusted to represent task-specific interaction.

GBA claims and measurement challenges

GBA design encourages us to consider reciprocal types of interactions, which raises fundamental measurement challenges. Next, we draw on the nascent research and development efforts in the GBA literature to highlight some measurement issues that demand thoughtful consideration and engaged research efforts. We address central measurement concepts such as generalizability, score comparability, and issues of dimensionality. A critical reader would likely term the claims “traditional” or “conventional” and would be warranted in such a critique. While such conventional claims likely “underrepresent the range, complexity, and diversity of the phenomena” (Moss et al. 2005, p. 64) related to dynamic assessment as embodied by GBAs, they reflect the current state of measurement discussions regarding GBAs. At this stage, we seek to problematize the issues for research purposes.

Generalizability

Claim 1: Performance on the task can be generalized to other like tasks in different contexts.

Claim 2: Systematic variation of task features controls construct irrelevant variance.

Tasks within GBAs are typically multifaceted situations, vary in degrees of difficulty as well as interactional complexity, and are used to measure a variety of skills. GBAs "allow us to make observations in contexts closer to those in the real world, creating complex scenarios required to evaluate the application of knowledge and skills" as well as present players with tasks that are "engaging and motivating" (DiCerbo 2014, p. 17). Our ability to generalize GBA task performance to like tasks is restricted by major factors such as a high level of contextualization and the default view of the nature of interaction.

Generalization of task performance is limited because tasks are highly contextualized, an imperative feature to promote motivation and engagement. To provide a specific example, we return to the previously cited research by DiCerbo (2014), where sufficient motivation is a requirement for the measurement of persistence. Motivation seems to be dependent on a myriad of elements including: task features, person characteristics, and cultural attitudes. These elements are a mix of both the intensive (i.e., task and person) and extensive context (i.e., cultural attitudes), which we contend is exactly where the assessment situation should be located. Chalhoub-Deville (2003) argues: "if internal attributes of ability are inextricably enmeshed with the specifics of a given situation, then any inferences about ability and performance in other contexts are questionable" (p. 376). She calls for research into the external contexts in which internal knowledge and processes are accessed in similar degrees and ways to allow for warranted generalization.

As of yet, research on GBAs has avoided systematic investigation of internal knowledge and processes and their connections to external domains. Current GBA best practices suggest systematic variation of game features to elicit a range of performances. DiCerbo et al. (2016) suggest that designers of GBAs recreate a task throughout the game space while systematically varying task features such as genre (collaborative, quest, etc.) and surface features such as setting (mountains, space, etc.) to promote varying degrees of motivation and engagement. The intent with these variations is to prevent any one task feature from dictating player interactions with the task. Implicit in these attempts to avoid a task effect is the prevalent view that tasks are not part of the construct of interest and variance apportioned to tasks is construct irrelevant variance.

Approaches that seek to vary task features effectively consider a reciprocal type of interaction, i.e., intertwined person and task, as a form of construct irrelevant variance to be controlled for, if it cannot be avoided. Chalhoub-Deville calls for "a shift from traditional examinations of the construct in terms of response consistency, to investigations that systematically explore inconsistent (which does not mean random) performances across contexts" (Chalhoub-Deville 2003, p. 378). She concludes that it is not appropriate to assume that variation in performance is not relevant to the construct of language use and interaction.

Score comparability

Claim 3: There are adequate opportunities to observe an individual's performance on measurement indicators.

Claim 4: Individuals do not need to be administered the same task to achieve score comparability.

Test taker/player freedom in GBAs can “decrease comparability of evidence across players” (Mislevy et al. 2015, p. 26). The amount of choice present in GBAs creates what Behrens and DiCerbo (2014) have termed a “digital ocean.” Players are given choices which create direction in their game play. Once executed, many decisions are irreversible and therefore have meaningful implications. The high stakes associated with decision-making in games potentially cultivates engagement and motivation. If unconsidered, the path players take through a game can also decrease the amount of evidence available from which to make inferences (DiCerbo et al. 2016).

Concerns regarding the quantity of evidence available to support inferences are answered by providing players with multiple opportunities to complete the same task while systematically varying the task features (DiCerbo et al. 2016). Varying task features eliminates redundancy and fatigue with the task. Multiple occasions to interact with the task ensure that regardless of player game path they will have exposure to at least some iterations of the task. Replications of task address not only concerns related to quantity of evidence but are to some extent necessary for addressing issues of reliability (Brennan 2001). In a discussion on reliability of performance assessments, Brennan concludes that reliability issues are due to “the combined effect of large person-task interactions *and* small numbers of tasks” (emphasis is original, p. 308). Guidelines for game design from DiCerbo et al. (2016) remedies the issue of small numbers of task. However, reliability would likely need to be redefined as we expect performance could substantially vary over tasks that have been designed to measure the same construct given the unique interaction between person and task each assessment situation would represent.

Davey et al. (2015) offer several methods for managing score comparability issues in performance assessments that are applicable to handling comparability issues in GBAs. The two methods of most relevance to GBA-related inferences are (1) “report only group-level scores,” and (2) “accept that scores are not fully comparable and limit inferences accordingly” (Davey et al. 2015, p. 52). Individual level scoring restricts generalization of scores to a narrowly defined universe of generalization. For a description of the universe of generalization, see Kane (2006). In a GBA scenario, it would suggest restricting comparability of scores to those players who received an identical “test form.” This would mean that scores may only be compared across individuals who took the exact same path through the game and were therefore administered identical tasks.

However, the issue of test form comparability is not foreign to the field of assessment. Computer adaptive testing (CAT) also results in forms which are uniquely constructed for individual examinees. With CAT, tasks are piloted to ascertain task difficulty and establish comparability of performance opportunities. With GBA, activity templates should be created and piloted to document how various combinations of task features interact with person characteristics and influence task difficulty (DiCerbo et al. 2016). However, given the number of activity templates that would need to be piloted, this may present an unreasonable burden on game designers.

Furthermore, as explicated in later discussions on multidimensionality, even when given identical tasks, individuals will bring a unique set of person characteristics, schooling experiences, cultural ideologies, etc. which will directly influence their interaction with the task. Consideration should be given to situations where group level scores are also appropriate. A group score can also contribute to a test taker's ultimate individual score.

Dimensionality

Claim 5: We can make comparable score and trait interpretations for individuals from different subpopulations.

Claim 6: We can parse out differences in performance due to the interaction of person characteristics and task features. In doing so, we can quantify the measurement error in scores due to confounding variables.

Concerns of *differential item functioning (DIF)* are attended to as part of score comparability. The *Standards for Educational and Psychological Testing* (AERA, APA, and NCME 2014) defines DIF as "a statistical indicator of the extent to which different groups of test takers who are at the same ability level have different frequencies of correct responses, or, in some cases, different rates of choosing various item options" (p. 218). DIF does not necessarily indicate item bias or flaws with the item. Rather, it could be indicative of multidimensionality in the test or differential impact, the case of differential performance on the item due to real subpopulation differences on the construct of interest.

To mitigate DIF, traditional assessments attempt to minimize construct irrelevant features. However, guiding principles for GBAs seek to vary the construct irrelevant variance inducing features over a larger number of tasks (DiCerbo et al. 2016). As such, there will inevitably be instances where those features create discrepancies in performances between subpopulations due to person by task interaction within the assessment situation. It is worth noting that these differences may not conform to traditional grouping covariates (e.g., gender, ethnicity, age), speaking to the need to expand upon traditional notions of person characteristics when examining non-traditional dynamic assessments such as GBAs. For example, does the player prefer collaborative tasks to individual tasks and therefore engage with those more making them appear easier?

Unidimensionality is operationalized as consistent variation across tasks, persons, and within proficiency levels. Example covariates which could be used to form a subpopulation include, but are not limited to: collaboration preference, level of motivation, gender, age group, and proficiency level. Examination of the construct within any *one* of these subpopulations may lead to a conclusion of unidimensionality. However, test takers do not typically fall into one category. Considering our discussion of intensive and extensive context, a multitude of factors can come to bear on an individual's task performance. The intersection of these factors leads to performances that can be characterized as multidimensional.

Latent class analysis may be a useful tool in forming groups based on similar response patterns and may address issues of multidimensionality. Finer groupings along with activity templates, discussed previously, allow us to tailor assessments, probe more accurately into performances, provide more appropriate scoring, and make better predictions. Key issues such as sample size within the finer groupings and the interpretability of groupings if it is not readily discernable what common underlying factors individuals within a group possess that are influencing their interaction with the task represent major challenges that would need to be addressed. Scoring and dimensionality would then be subgroup dependent.

Not all person characteristics will be of interest. Some, such as familiarity with or access to computers, are simply confounding variables. A difficulty posed by reciprocal interaction is separating out the variability caused by person-by-task interactions of interest versus confounding person-by-task interactions. In essence, determining what is multidimensionality and what is DIF. Current literature on GBA offers no considerations on DIF analyses. Not only do we need to investigate issues of score comparability across different task paths but also across subpopulations.

Conclusions

The nature of the construct in language assessment has generally been assumed to be a reasonably stable entity that resides within the person. Conducting assessment under this assumption allows test professionals to detach student performance from tasks and develop inferences regarding student ability (Deville and Chalhoub-Deville 2006). This assumption "is a cornerstone when discussing traditional notions of reliability and validity" (Deville and Chalhoub-Deville 2006, p. 12) and is in line with interdependent interaction language and measurement frameworks, which dominate the language testing and traditional measurement literature.

In the article, we push for moving away from interdependent constructs to consideration of interaction as reciprocal. This is especially necessary, we argue, when working with innovative dynamic assessment systems such as GBA. GBAs represent exciting research and development endeavors that can allow explorations of students' capabilities in a variety of complex and real-life domains. GBAs portray interactions that represent the interconnectedness of tasks and persons as a performative system in specific contexts. GBA interactions motivate us to consider person and task variables as inseparable and constantly working to change one another not only in terms of an intensive but also an extensive context.

Concepts taken from ability-in language user-in context, such as the entanglement of person and task and role of extensive context in shaping performance, lend themselves particularly well to dynamic GBAs. Such concepts, nevertheless, require reexamination of fundamental measurement and validation principles. The issues we have touched on in the areas of generalizability, score comparability, and dimensionality, provide examples of where investigations are needed.

While only a handful of the Asian game-based learning literature reviewed for this paper has been formally discussed, it is interesting to note eight of nine articles surveyed were supported with government funding (in addition to those previously cited see also Hooshyar et al. 2016; Chuang et al. 2015; Wang et al. 2016; Hsu, Liang, and Su, 2015; Cheng et al. 2012; Liu and Chu 2010). This support tends to speak perhaps to national interests and/or availability of funds to promote this area of research and development. The language testing community in Asia, however, does not seem to be as engaged in this research and development endeavor. (The articles reviewed are limited to journals dedicated to education and technology.) We invite language testers in general, and those in the Asian language testing community in particular, to engage in GBAs and to undertake the transformative design and research needed to accommodate the changing assessment and measurement realities.

Acknowledgements
Not applicable.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Authors' contributions
All authors contributed to and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 29 March 2017 Accepted: 8 September 2017
Published online: 09 October 2017

References
American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests* (Vol. 1). Oxford: Oxford University Press.

Behrens, J. T., & DiCerbo, K. E. (2014). Harnessing the currents of the digital ocean. In J. A. Larusson & B. White (Eds.), *Learning analytics: from research to practice* (pp. 39–60). New York: Springer.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*(4), 295–317.

Bronfenbrenner, U. (1979). The ecology of human development: Experiments by nature and design. Cambridge, MA: Harvard University Press.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20*, 369–383.

Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In R. Lissitz (Ed.), *The concept of validity: revisions, new directions and applications* (pp. 65–81). Charlotte: Information Age Publishing Inc..

Chalhoub-Deville, M., & Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. Brennan (Ed.), *Educational measurement* (pp. 517–530). Westport: American Council on Education and Praeger Publishers.

Chapelle, C. (1988). Construct definition and validity inquiry in SLA research. In L.F. Bachman and A.D. Cohen (eds.), Interfaces between Second Language Acquisition and Language Testing Research. New York: Cambridge University Press, pp. 32–70.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Cheng, Y. M., Kuo, S. H., Lou, S. J., & Shih, R. C. (2012). The construction of an online competitive game-based learning system for junior high school students. *The Turkish Online Journal of Educational Technology, 12*, 2.

Chiu, Y. H., Kao, C. W., & Reynolds, B. L. (2012). The relative effectiveness of digital games-based learning types in English as a foreign language setting: a meta-analysis. *British Journal of Educational Technology, 43*, 104–107.

Chuang, T. Y., Liu, E. Z. F., & Shiu, W. Y. (2015). Game-based creativity assessment system: the application of fuzzy theory. *Multimedia Tools Applications, 74*, 9141–9155.

Craig, A. B., DeRosier, M. E., & Watanabe, Y. (2015). Differences between Japanese and US children's performance on "zoo U": a game based social skills assessment. *Games for Health Journal: Research, Development, and Clinical Applications, 4*(4), 285–294.

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment*. Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service.

Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability. *Inference and generalizability in applied linguistics: Multiple perspectives, 12*, 9.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Educational Technology & Society, 17*(1), 17–28.

DiCerbo, K. E., Mislevy, R. J., & Behrens, J. T. (2016). Inference in game-based assessment. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment* (pp. 253–279). New York: Routledge.

Flowers, C., Turner, C., Herrera, B., Towles-Reeves, L., Thurlow, M., Davidson, A., & Hagge, S. (2015). *Developing a large-scale assessment using components of evidence-centered design: did it work?* Chicago: Paper presented at the annual National Council on Measurement in Education.

Gilbert, R. (1992). Text and context in qualitative educational research: Discourse analysis and the problem of contextual explanation. *Linguistics and Education, 4*(1), 37–57.

Hao, J., Liu, L., von Davier, A., & Kyllonen, P. (2015). Assessing collaborative problem solving with simulation based tasks. In O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the material conditions of learning: computer supported collaborative learning*. Gothenburg: Proceedings of 11th international conference on computer supported collaborative learning. https://www.isls.org/cscl2015/papers/MC-0297-ShortPaper-Hao.pdf.

Haywood, H. C., Tzuriel, D., & Vaught, S. (1992). Psychoeducational assessment from a transactional perspective. In *Interactive assessment* (pp. 38–63). New York: Springer.

Hooshyar, D., Ahmad, R. B., Yousefi, M., Fathi, M., Abdollahi, A., & Horng, S. J. (2016). A solution based intelligent tutoring system integrated with an online game based formative assessment: development and evaluation. *Education Tech Research. Development, 64*, 787–808.

Hsu, C. Y., Liang, J. C., & Su, Y. C. (2015). The role of the TPACK in game-based teaching: does instructional sequence matter? *Asia Pacific. Educational Researcher, 24*, 463–470.

Hung, H. C., Young, S. S., & Lin, C. P. (2015). No student left behind: a collaborative and competitive game-based learning environment to reduce the achievement gap of EFL students in Taiwan. *Technology, Pedagogy and Education, 24*, 35–49.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and. Practice, 18*, 5–17.

Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulations. CRESST Report 771*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Krahnke, K. (1987). *Approaches to syllabus design for foreign language teaching*. Englewood: Prentice-Hall, Inc..

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70*, 366–372.

Kramsch, C. (1998). *Language and culture*. Oxford: Oxford University Press.

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

Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly, 26*, 27–56.

Lopez, A. A., Guzman-Orth, D., & Turkan, S. (in press). The use of translanguaging to assess the mathematics knowledge of emergent bilingual students. *Language Assessment Quarterly*.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: foundations, innovations, and perspectives* (pp. 59–81). New York: Springer.

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., & John, M. (2015). Psychometrics and game-based assessment. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 23–48). New York: Routledge.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and. Perspectives, 1*, 3–67.

Moss, P., Pullin, D., Gee, J.P, & Haertel, E.H. (2005). The idea of testing: psychometric and sociocultural perspectives. Measurement Interdisciplinary Research and Perspectives 3(2): 63–83.

Nelson, B. C., Erlandson, B., & Denham, A. (2011). Global channels of evidence for learning and assessment in complex game environments. *British Journal of Educational Technology, 42*, 88–100.

Perez, R. S., Ralph, J., & Niehaus, J. (2016). The role of neurobiology in teaching and assessing games. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment* (pp. 163–179). New York: Routledge.

Popp, E. C., Tuzinski, K., & Fetzer, M. (2016). Actor or avatar? Considerations for selecting appropriate formats for assessment content. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 79–103). New York: Routledge.

Purpura, J. E. (2008). Assessing communicative language ability: models and their components. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education, Vol. 7. Language testing and assessment* (2nd ed., pp. 53–68). Dordrecht: Kluwer.

Ranalli, J. (2008). Learning English with the Sims: exploiting authentic computer simulation games for L2 learning. *Computer Assisted Language Learning, 21*(5), 441–455.

Rankin, Y. A., Gold, R., & Gooch, B. (2006, September). 3D role-playing games as language learning tools. In Eurographics (Education Papers) (pp. 33–38).

Rankin, Y. A., McNeal, M., Shute, M. W., & Gooch, B. (2008). User centered game design: evaluating massive multiplayer online role playing games for second language acquisition. In *Proceedings of the 2008 ACM SIGGRAPH symposium on video games* (pp. 43–49). ACM.

Rueda, R., O'Neil, H. F., & Son, E. (2016). The role of motivation, affect, and engagement in simulation/game environments: A proposed model. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment* (pp. 203–229). New York: Routledge.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment, 8*(4), 4–47.

Schwabe, F., von Davier, A. A., & Chahoub-Deville, M. (2016). Language and culture in testing. *The ITC International Handbook of Testing and Assessment*, 300–317.

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: foundations, innovations, and perspectives* (pp. 59–81). New York: Springer.

Sireci, S. G., & Faulkner-Bond, M. (2016). The times they are a-changing but the song remains the same: future issues and practices in test validation. In C. Wells, M. Faulkner-Bond, & E. Hambleton (Eds.), *Educational measurement: From foundations to future* (pp. 435–448). New York: Guilford Press.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context* (pp. 3–37). Cambridge: Cambridge University Press.

Tzuriel, D. (2001). Dynamic-interactive approaches to assessment of learning potential. In *Dynamic assessment of young children* (pp. 1–10). New York: Kluwer Academic/Plenum Publishers.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior, 29*, 2568–2572.

Wang, J. H., Chen, S. Y., Chang, B., & Chan, T. W. (2016). From integrative to game-based integrative peer response: High ability versus low ability. *Journal of Computer Assisted Learning, 32*, 170–185.