sciendo
ExELL (Explorations in English Language and Linguistics)
5.1 (2017): 55-70
DOI: 10.2478/exell-2018-0004
Original scientific article

Examining validity in computerized dynamic assessment

Zaha Alonazi
Iowa State University, USA
Majmaah University, KSA

Abstract

Computerized dynamic assessment (CDA) posits itself as a new type of assessment that includes mediation in the assessment process. Proponents of dynamic assessment (DA) in general and CDA in particular argue that the goals of DA are in congruence with the concept of validity that underscores the social consequences of test use and the integration of learning and assessment (Sternberg & Grigorenko, 2002; Poehner, 2008; Shabani, 2012;). However, empirical research on CDA falls short in supporting such an argument. Empirical studies on CDA are riddled with ill- defined constructs and insufficient supporting evidence in regard to the aspects of validity postulated by Messick (1989, 1990, 1996). Due to the scarcity of research on CDA, this paper explores the potentials and the viability of this intervention-based assessment in computer assisted language testing context in light of its conformity with Messick's unitary view of validity. The paper begins with a discussion of the theoretical foundations and models of DA. It then proceeds to discuss the differences between DA and non-dynamic assessment (NDA) measures before critically appraising the empirical studies on CDA. The critical review of the findings in CDA literature aims at shedding light on some drawbacks in the design of CDA research and the compatibility of the concept of construct validity in CDA with Messick's (1989) unitary concept of validity. The review of CDA concludes with some recommendations for rectifying gaps to establish CDA in a more prominent position in computerized language testing.

Key words: dynamic assessment; computerized dynamic assessment; construct validity; sequential validity.

1. Introduction

While underscoring the need for integrating learning and assessment, McNamara (2001) criticized current approaches of assessment for prioritizing institutional needs over those of teachers and learners. He stressed that placing learners and teachers' needs secondary in the assessment process

can result in both theoretically and empirically underestimating the needs of learners and teachers. Unfortunately, although dynamic assessment (DA) has been presented by its proponents as a new generation type of assessment that prioritizes learning, insufficient research has been done in L2 and studies examining its potentials in computer-based environment are scarce. Even worse, the majority of the existing studies on this issue are riddled with major deficiencies, such as neglecting the discussion on the construct validity for test use and the meaningfulness of the scores. Advocates of DA contend that the procedures and goals of DA conform to Messick's (1989) concept of validity which subsumes both actual and potential consequences of a test use. Nevertheless, as to be discussed in coming sections, the actual research on DA was not successful in conceptualizing the holistic view of the validity it claims to embrace.

Given the increasing interest in utilizing computers in second language assessment, this paper aims at providing a comprehensive evaluative judgment of the actual potentials and challenges of computerized dynamic assessment (CDA) in the context of second language testing. Serving this purpose, the discussion in this paper will begin with an overview of the theoretical orientations, tenets and models of DA. It will then proceed to evaluate the general research design of CDA studies with particular emphasis on the construct validity and its claimed conformity with Messick's holistic view of validity. Finally, the paper will conclude with a discussion and an analysis of the drawn conclusions in an attempt to envision the potentials of CDA and the challenges it faces.

2. What is dynamic assessment?

While teaching for testing has been described by the literature of assessment as a negative washback of assessment which should be avoided, dynamic assessment has been proposed as a means for integrating learning and assessment, a legitimate goal for assessment research. The use of DA came in response to the criticism of the goals and procedures of conventional or standardized tests which perceive human abilities as discrete, stable entities that can be measured quantitatively (Ratner, 1997). Dynamic assessment, as the name suggests, is an approach that integrates both assessment and instruction in its procedures aiming at improving and disclosing the learner's current and potential learning abilities within the framework of a zone of proximal development (ZPD) (Sternberg & Grigorenko, 2002; Poehner, 2008). The zone of proximal development refers to "the distance between the actual development level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers"

(Vygotsky, 1978: 86). To Vygotsky, learning starts at the intermental or social level before moving to the intramental level, i.e. becoming individualized. As such, he differentiates between two types of learning: learning with assistance and independent learning. The second stage depends largely on the first one. Mediation in this sense is fundamental for cognitive development and for the ultimate goal of independent learning. Through meditation, one can capture the whole picture of learners' cognitive development through assessing not only current but also potential improvements that are in the process of maturation (Vygotsky, 1978: 87). Sternberg & Grigorenko summed up the major differences between DA and non-dynamic tests as follows:

> In this form of testing, each examinee receives one or more items, as in static testing. But rather than scores being based simply on performance on the initial presentation of these items, the score is based on a system that takes into account the results of an intervention. In this intervention, the examiner teaches the examinee how to perform better on individual items or on the test as a whole. The final score may be the learning score representing the difference between pretest (before learning) and posttest (after learning) scores, or it may be the score on the posttest considered alone (Sternberg & Grigorenko 2002: vii).

Advocates of DA argue that although standardized tests are reliable for testing physical phenomena that follow a systematic pattern of development, they fall short of being able to describe constantly changing cognitive abilities (Feuerstein, Rand & Rynders, 1988; Lantolf & Thorne, 2006; Poehner, 2008). Drawing upon this nature of human cognitive functioning as well as Vygotsky's theories about the role of social interaction in language development, DA takes mediation as a cornerstone for valid and reliable assessment of cognitive abilities. McNamara argued that proficiency should not be perceived as a discrete ability that can be measured in a "curious kind of isolation"(1997: 449), but rather as a developing and changing phenomenon; hence, the use of assistance should be considered as a valid measure for anticipating learners' growth potentials. As such, McNamara (2001) accentuated the need to include systematic sustained reflection from either the teacher or the learner on the quality of learners' products as part of the assessment process. He stressed that such reflection should not be restricted to the traditional comparative based assessment but to any critical type of reflection that aims at raising learners' awareness of their learning development. DA is, however, not only concerned with the improvement during the mediation process, but also in transferring the cognitive development to novel, relevant tasks (Poehner, 2007). Based on the type of mediation and the procedures followed, two types of mediation are discussed in DA research.

Two major approaches are described in DA literature: interventionist and interactionist. According to Poehner (2008: 18), mediation refers to any type of assistance ranging from "standardized hints to dialogic interaction". In

the interventionist approach, as the name suggests, mediation is provided based on a diagnostic analysis of learners' needs as illuminated by their performance on a pretest. The goal is to accelerate the rate of development utilizing a prescribed set of sequenced prompts or mediation types. Therefore, they are more oriented toward the standardized procedures of administration (pre and posttest) in psychometric test measures. On the other hand, mediation in the interactionist approach is cyclic and is fine-tuned toward learners' emerging and individualized needs. In line with these approaches, two main types of DA are discussed in the literature: Sandwich and Cake formats as termed by Sternberg & Grigorenko (2002). The former is more in line with interventionists' approach as the treatment or mediation is "sandwiched" between the pre and posttest. The latter is like a cake with multiple layers of mediation based on learners' emerging needs. Due to its quantitative orientation, the Sandwich type is more appropriate with a large number of individuals whereas the latter is more viable with a small number of participants since it is more qualitatively oriented (Lantolf & Poehner, 2004).

3. Dynamic assessment vs. non-dynamic assessment

Although some researchers categorized tests into either static (referring to the conventional type of testing) and dynamic (Sternberg & Grigorenko, 2002), Poehner (2008) rejected such a division, preferring the use of DA and non-DA instead on the ground that the so called static tests may encompass other types of assessment with dynamic practices, such as assessments where feedback is provided, e.g. a portfolio. Hence, in his view the two types should be conceived in a continuum rather than being dichotomous.

By and large, four main differences should be considered between DA and non-DA: the goal or assumptions of DA versus non-DA, the role of examiners or assessors, the type of feedback provided (Sternberg & Grigorenko, 2002), and finally the concept of the test's validity. Conventional tests perceive cognitive abilities as static, fixed entities that could be quantified using psychometric type of testing. This notion is completely rejected in DA where cognitive abilities are perceived as a moving target or of a modifiable nature that emerge with appropriately tailored assistance. As such, the purpose of DA is to identify what learners are developmentally capable of achieving and what cognitive functions are involved in the process of maturation (Poehner, 2008). Hence, in congruence with Vygotsky's proposition, the goal of DA allows for noticing signs of emerging development, and paves a way for a finely tuned intervention. Therefore, non-dynamic tests are accused of embracing a past to present perspective where learning capabilities are considered to remain stable across a period of time whereas DA takes a present to future view of cognitive potentials that learners may start

at a "zero point" of development in the skills to be tested and with assistance can move to a higher developmental level (Sternberg & Grigorenko, 2002: 29). Another major difference between dynamic and non-dynamic assessment lies in the nature of the relationship between the assessor and the examinee. In DA, where mediation is perceived as a prerequisite for cognitive or language development, assessors take the role of mediators or the assistance providers whereas such a practice is considered to be a threat to the consistency of results and the ethics of assessment in conventional types of testing (Sternberg & Grigorenko, 2002).

Predictive validity is also viewed differently by DA researchers. In conventional testing, predictive validity is considered established when there is a positive, moderate or strong correlation between scores and real-life tasks, e.g. academic performance (Porte, 2010). On the other hand, in agreement with Ratner (1997), Poehner (2008) believes that a strong or weak correlation between two sets of scores measuring the same construct does not necessarily imply the validity of either instrument because the same construct may have various representations in different settings. Poehner argues that a major difference between conventional tests and DA is the quality of prediction, contending that unlike in non-dynamic tests, the examinee's performance in DA should not be considered linear provided that intervention can change the course of cognitive development during the test performance.

With this in mind and in an attempt to provide a robust model of validity that accounts for teaching and assessment, Poehner (2011) proposed a framework of validity, in response to Moss's call for a validity "framework to guide thinking and actions" (2003: 15). He introduced two types of validity for DA: micro and macro validity. The first is analogous to a test's single item validity since it focuses on the mediator's moves and the consequent interpretation of learners' abilities with the supporting evidence of such interpretation. The macro level of validity corresponds to the validity of the whole test. As in conventional tests, it is concerned with the validity of interpretation of the whole DA session, i.e. the evidence of cognitive development as a result of mediation. Particularly, it considers three major components: the quality of mediation in terms of the moves used by the mediator, the corresponding changes in learners' responses, and the learners' verbalization of knowledge.

Although Poehner's model seems to be an actual attempt to establish the validity of DA, there is no evidence that such a notion is applicable to CDA. Even if so, the proposed model of validity is of a subjective nature that is more conceptually rather than practically established. It, in my view, tends to raise more questions than it answers. For example, what determines the quality of mediation or the moves? What qualifies for appropriate response? What qualifies for a transfer of knowledge in learners' responses? More im-

portantly, how can the micro and macro level of validity be established quantitatively? How can the correlation between moves and responses be established if such a model is to be transferred or adopted into a computerized version? And how can this model account for situations where learners provide correct responses, but are unable to explain or reason their choices? Finally, the model gives no indication to construct validity which Poehner (2008, 2011) himself has emphasized repeatedly.

At any rate, taking cognitive development as a major criterion for test validity, DA has been posited by its proponents to be in line with Messick's (1989) validity proposition. Poehner succinctly stated this: "DA represents one response to Messick's concern, as learner development becomes the immediate consequence, and indeed the primary goal of the procedure" (2008: 76). Consequential validity, as the name suggests, accentuates the meaning of the scores as well as the social consequences of use in a particular context as one important facet of validity (Messick, 1989). In his view of test validation, Messick (1990, 1996) proposed a unitary view of validity, in which construct validity is broadened to subsume empirical and theoretical evidence of six facets of validity, including consequential validity. The six aspects of construct validity include content, substantive, structural, generalizability, external and consequential aspects of test use and interpretation (Messick, 1996). Content relevance refers to the test content relevance to the larger domain in question, whereas substantiveness indicates "the extent to which the context of the items included in (and excluded from) the test can be accounted for in terms of the trait believed to be measured and the context of measurement" (Loevinger, 1957: 661). The structural aspect of construct validity refers to the internal structure of the test, i.e. the consistency of the test's tasks with those of the construct domain (Messick, 1989, 1996). Generalizability is another aspect of construct validity meant to ensure that the test score interpretation is not limited to the test *per se* but extends to the larger focal construct. Hence it requires evidence of performance consistency across tasks pertinent to the broader construct (Messick, 1996). As such, the major concern is not the generalizability across different populations or settings, since some constructs, e.g. mood, are vulnerable to change overtime, but rather it is the nature of the construct in question and its "theoretical applicability" (Messick, 1980: 1019). External aspect or criterion relatedness refers to the correlation between the test scores and other measures of the same or theoretically related constructs. According to Messick (1996: 251), the most important external relationships are those with criterion measures "pertinent to selection, placement, licensure, certification of competence, program evaluation, or other accountability purposes in applied settings."

The consequential aspect of construct validity requires appraisal of both the intended and unintended consequences of test use and interpretation in

both the short and the long term (Messick, 1996). Washback, which refers to the positive or negative learning and teaching effects associated with test use, has particularly been emphasized as an important form of the consequential aspect of construct validity (Messick, 1996). Messick, however, warned against viewing consequential validity as one independent form of validity rejecting the compartmenting of the various aspects of validity. He argued that such a view gives the illusion that one type of validity can compensate for the others. Validation of test-based inferences requires multiple types of evidence rather than multiple types of validity (Messick, 1980).

Messick (1989) pinpointed that what basically renders a test valid or invalid is not the positive or the negative consequences *per se*, but the fact that these consequences are traced directly to the construct of the test. Such coherent perception was also reiterated by Frederiksen & Collins (1989) when they emphasized that validity is dependent on the specifications of the construct that the test intends to measure because the use of the test can induce either positive or negative changes at both educational and societal levels. As such, construct under representation and construct irrelevant variance have been postulated as major threats to construct validity. The former indicates that the test's content or tasks are not sufficiently reflective of the characteristics of the target domain construct. The latter means that some portions of the test score variance that accounts for the observed performance is attributed to factors irrelevant to the target construct such as fatigue, low motivation or repeated practice, lack of familiarity with the test, etc. (Messick, 1989, 1996). Hence, ensuring that the reported washback is pertinent to the test use requires not only appropriate representation of the test construct but also minimization of the effects of any potential construct irrelevant factors. With this in mind, validation of a test use can be attained through validating test design which in turn acts as a basis for the evaluation of washback (Messick, 1996).

By and large, the claim that DA's concept of validity conforms to the one postulated by Messick falls short of being empirically evident. Since the discussion of the research on DA in second language assessment is beyond the scope and the purpose of this paper, the critical review will be limited to the use of computerized versions of DA (CDA) in second language assessment. The literature on CDA in second language assessment, as will be shortly discussed in the coming section, shows little evidence of conformity with Messick's view of validity. It suffers from some shortcomings in methodological design pertinent mainly but not exclusively to the lack of evidence for the various aspects of construct validity, namely, (a) the specification of the construct to be measured and its relevance to the targeted domain, and (b) evidential basis for CDA utility and appropriateness for the consequent decision- making processes.

4. Critical review of CDA empirical research in L2 assessment

Studies on CDA in second language acquisition (SLA) are still limited. However, they cover a wide range of language skills, as will be discussed later in the section. To evaluate CDA's contribution to second language testing and learners' development, the forthcoming discussion provides a critical review of the literature on CDA in a second language (SL) context highlighting the major findings of CDA research and appraising its level of compatibility with the unitary view of validity construct with a particular focus on consequential validity as posited by Messick (1989). To the author's knowledge, the reviewed studies constitute the only studies conducted on CDA in second language learning contexts. Serving the purpose of this paper, the inclusion criteria were set to include empirical papers on CDA in second language settings. Terms such as *computer and DA*, and *DA in second language* were used to search major databases such as Google Scholar and Academic Search Premier (EBSCO). As such the study by Shrestha & Coffin (2012) was excluded from the review since the purpose of their study does not seem to target second language learners in particular, given that one of the two participants was a native speaker of the target language.

It is noteworthy that despite their limited number, the studies on CDA dealt with various language skills including reading comprehension (Shabani, 2012; Teo & Jen, 2012) reading and listening (Poehner & Lantolf, 2013), vocabulary (Jacobs, 2001), listening and speaking (Lin, 2010) as well as writing (Birjandi & Ebadi, 2012). Nevertheless, the research on CDA shares the commonality of obscurity of the terms used and the ill-defined constructs to be measured, as will be illustrated in the paragraphs below.

Only one study reported the use of CDA for assessing vocabulary knowledge. Jacobs (2001) reported the use of a modified version of an interactive program called KIDTALK based on DA principles. The program aimed at assessing preschool children's language aptitude through teaching an invented Swahili based language presented by puppets in videos. After the first training sessions, learners who came from ethnically and culturally different backgrounds were treated by CDA through KIDTALK. At the end of assessment, the computer generated two reports for each student. One included only the correct answers regardless of the number of prompts used and the other gave a more detailed description of the type of mediation and failed attempts of the learners.

While the study revealed some important details about the subjects who had difficulty arriving at the correct answer, a finding that may not be revealed by traditional tests, evaluation of the construct validity of the test was not possible. The supposedly measured constructs of "language learnability" and "potential abilities" were not clearly defined. Additionally, no replica-

tion or follow-up study was reported to validate the findings and the actual consequences. Given the ill-defined constructs, attempts of replication would be difficult.

Two studies reported positive effects of using CDA for improving reading skills (Shabani, 2012; Teo & Jen, 2012). Shabani (2012) examined the effectiveness of CDA in improving EFL college students' reading comprehension by comparing learners' performance on CDA and computerized non-dynamic assessment (CNDA). It was found that in CNDA, 79% of students were labelled as "non-gainers" whereas 88% were labelled as "gainers" by the CDA. Shabani stressed that with CNDA, learners' potential zone of development and differences in their ZPDs would remain undisclosed. He concluded that the results of CDA were helpful in making placement and selection decisions.

Although Shabani's study aimed to examine the discriminant validity of CDA, there is no substantive evidence that justifies the subsequent proclaimed decisions as the study gave no indication of the type of reading skills or the boundaries of the construct of reading comprehension to be assessed or even the nature of tasks used in the test. Similarly, the terms "gainers" and "non-gainers" and the classification of learners' ZPD levels are used ambiguously because the scores' interpretation underlying such a division were not clearly defined. Although Shabani concluded that validity in DA in agreement with Messick is reflected by the change brought about by the treatment, we would argue that validity in Messick's view is a unitary concept in which the various complementary facets of validity evidence need to be substantiated empirically and theoretically - a condition that was not manifested in the current study.

Reporting similar improvement in inferential reading skills after an eight-week CDA treatment, Teo & Jen's (2012) study is not however, without methodological deficits. Compared to other studies on CDA, their study can be perceived to some extent as well designed since the researchers specified the boundaries of the construct to be assessed and the validity of the assessment instrument used. Moreover, the study provided a detailed explanation of the type and levels of mediation in their test supported with examples. One drawback, however, is that they attempted to generalize the interpretation of learners' metacognitive reading strategies used in the test; they conducted no statistical analysis on the type or the rate of improvement nor on the percentage of learners who manifested metacognitive awareness of the reading processes strategies.

The only study reporting improvement in speaking and listening using CDA was that of Lin (2010) in which both interactionist and interventionists' approaches in CDA were used for kindergarten EFL learners. Tailored to learners' age and needs, three types of mediation were used in a 20 week-

intervention: repetition, use of L1 and non-verbal cues with the content of tasks being representative to daily L2 use. The results showed that children improved in six out of the seven assigned tasks. It was concluded that "interactive DA is a desirable alternative to NDA in the current EFL education practices with young EFL learners" (Lin, 2010: 286). Lin's study, however, is not without flaws. Lin indicated that the study was part of an intervention program at an urban kindergarten in China. However, no information was provided on the nature and goals of the other parts of the program, which makes it difficult to attribute the findings solely to the intervention in question.

Birjandi & Ebadi (2012) explored the impact of mediation on the assessment of two EFL female learners' use of modals. The types of mediation ranged from implicit mediation represented by various degrees of textual comments to explicit mediation that employed live Skype conferences. An analysis of the quality and frequency of mediation using Aljaafreh's & Lantolf's (1994) five transitional levels of mediation revealed that assessment helped learners progress from other-regulated to self-regulated learning but with different degrees. Similar to Shabani's (2012) findings, they concluded that although both participants would have been labelled as non-gainers by traditional tests, the assessment revealed differences in their understanding of the target form.

The small number of participants and the qualitative and subjective nature of analysis of Birjandi & Ebadi (2012) makes it difficult to replicate their study and to evaluate the validity of conclusions. Moreover, there was no indication of how the researchers in this study addressed the construct irrelevant variances since we were not told whether the participants were enrolled in other writing courses nor was there any indication of the participants' writing proficiency level at the outset of the study. The researchers claimed that learners' performance on transcendence activities revealed "microgenetic development" in writing, and yet no evidence was provided to support such a claim – not to mention that the construct of "microgenetic development" was not defined.

In an attempt to introduce CDA as a quantifiable assessment approach for a large number of EFL learners, Poehner & Lantolf (2013) explored the potentials of CDA in improving listening and reading comprehension skills among learners of French, Chinese and Russian. To identify learners' current and potential learning abilities, three scores were generated: scores of independent unassisted performance, scores of assisted performance and a Learning Potential Score (LPS) which represented the difference between mediated and unmediated performance. To test learners' internalization of mediation, more difficult items were integrated into the original task and were scored separately. It was found that students with low performance on

unmediated tasks, performed high on assisted tasks. It was also found that some students with high LPS had "reasonable" scores on transfer tasks whereas others with low LPS performed worse on the same tasks. The study called for utilizing CDA for placement purposes and for predicting learners' responses to future instruction as those with low LPS seemed to require a more intensive type of intervention. They put, "We see LPS as potentially quite relevant to placement decisions whereby learners receive instruction that is complementary not to their level of actual development but to their level of proximal development" (Poehner & Lantolf, 2013: 337).

Similar to the majority of CDA empirical research, Poehner & Lantolf's definitions of the proposed constructs of the reading and listening comprehension tests are vague and broad. The constructs were operationalized as: "lexis, grammar, discourse and culture" (Poehner & Lantolf, 2013: 330) for the reading comprehension test and phonology for the listening comprehension test. As such, the reader is left with several unanswered questions. For example: how can the scores be interpreted in relevance to the assessed construct and what relevance do they pertain to the real-life use of the target language? What does "difficulty" involve as a criterion for selecting transfer items? Which particular aspect of the construct is assessed for the transfer of knowledge?

This discussion of the CDA empirical research illustrated the insufficient inferential and consequential basis to support the claim of its utility in second language assessment. Some of the above studies argued for the usefulness of CDA for admission and placement purposes. However, as Messick (1996) indicated, ascertaining score based actions particularly of high or medium stakes requires not only short term but also long term evidence, a condition not met in the previous studies. Even more importantly, none of the aforementioned studies attempted to define the construct of language proficiency which may undermine the criteria to be used for the suggested selection purpose. Consequently, it is not clear how the scores can be meaningfully translated to indicate learners' language proficiency level; what exactly the terms "gainers" and "non-gainers" refer to in terms of the assessed construct. In other words, what determines substantial from unsubstantial gain and on what basis is such a degree of gain determined? In fact, none of the discussed studies but Teo & Jen's (2012) attempted to clearly define either the target construct or the notion of the presupposed development. The threats to construct validity, i.e. construct under representation and construct irrelevant variance (Messick, 1990), were not appropriately, if at all addressed, which raises questions on the meaningfulness and the utility of the scores' interpretation and proposed use. In addition, all the studies discussed above adopted a within group design as no control groups were reported which to some extent makes it difficult to confidently conclude that

the observed performance solely resulted from CDA treatment rather than from repeated practice.

By and large, the flaws and fuzziness surrounding the construct validity, the interpretation and use of scores have made CDA less appealing in computer assisted language testing (CALT) settings which is to some extent reflected by the scarcity of research on CDA in second language assessment. This might be understandable in light of the role construct validity plays in evaluating the usefulness, the consequences and the contribution of a particular test to our understanding of the nature of second language learning. Emphasizing the role of test validity in second language assessment, Chapelle (2003: 172) argued that "If technology-based tests were accepted or dismissed without considering their validity, no progress will be made in SLA research". In fact, calling upon CDA proponents to clearly define the construct and provide sufficient evidence of the construct validity, which includes the consequent uses neither implicates, narrowing the focus down, nor discards the underlying principles of such a new generation assessment. Instead, it cements the claim of conformity with Messick's view of validity and allows for future replication studies to validate the findings.

5. Future directions and suggested studies

The literature of CDA reveals not only several methodological flaws but also some important untapped issues. On the one hand, the literature on CDA lacks research comparing CDA versions with paper-based versions and whether the change of modality affects the prospective gains of mediation. On the other hand, there is insufficient evidence on how CDA and CNDA differ in their interpretation of scores for a particular construct and how such difference affects the use of the test in question. Research on CDA has focused on assisting learners to do what they were unable to accomplish alone, but there has been no exploration of what causes such difficulties. It would be insightful to assemble such information from learners' own reflections and self-assessment procedures. In this vein, it would also be informative to examine the extent of involvement with the tasks and the type of test strategies that learners utilize in CDA vs. NCDA. Another aspect that has been overlooked by research on CDA is whether certain types of mediation would be more appropriate to certain language proficiency levels than others and how mediation-based assessment contributes to learners' self-confidence and hence motivation in second language learning.

There is also a shortage of empirical support for the long-term expected positive or even negative washback of CDA on learners' performance, instructional approaches and the overall impact on designing language pro-

grams. Follow-up studies that tackle the implications of CDA on all involved stakeholders is significant for consolidating the validity and reliability of such assessments in a second language learning context. We pass Messick' call (1989) for establishing consequential validity as an integral aspect of construct validity to CDA researchers: How can CDA procedures in contrast to NCDA promote equality and fairness in decision-making processes for learners with disadvantages either as a result of their social status or diverse background knowledge? The need for longitudinal studies in CALT was reiterated by Chapelle & Douglas (2006: 5): "The complexity inherent in computer–assisted diagnostic assessment calls for a sustained research agenda rather than a one-time project."

Nevertheless, and despite the various limitations of CDA, one cannot overlook its potentials for revolutionizing the concepts of assessment in SL given the centrality of learners' development to CDA, which is strongly accentuated in recent concepts of validity. A quick look at the previously discussed studies yields various important conclusions. CDA allows for practical and to some extent standardized implementations of principles established for DA in SLA, as both individualized and standardized types of mediation are viable given CALL affordances. In fact, with technology, response analysis and recordkeeping become more feasible than with traditional types of testing. Mediated interaction and informative feedback has not only unveiled what otherwise would be undisclosed cognitive potentials but could also contribute to learning autonomy when feedback is tailored to individual needs - an ultimate goal in all language learning settings. Given the affordances of technology in generating detailed records of learners' responses, CDA has the potential of being used as a formative assessment tool to help learners improve their language skills, to assist teachers in pinpointing where exactly individual learners need assistance and to accordingly design intensive intervention programs calibrated to learners' needs.

6. Conclusion

One expected goal of computer-based testing is to enhance learners' performance and assess learning abilities that traditional testing may fall short of disclosing. Advocates of DA and its computerized versions label it as a new generation of assessments that integrate learning and assessment and hence prioritize learning development through unveiling what otherwise might be latent cognitive abilities. With such goals in mind, DA and CDA in particular have been presented to be in line with the validity concept that takes washback or consequences of test score interpretation and use into consideration. However, as was concluded from the above discussion, the empirical evidence of such an argument is still defective. The fuzziness of the assessed

constructs and the lack of complementary evidence from the various aspects of validity as set by Messick (1996) weaken the argument that CDA is in line with the view of validity that emphasizes consequential validity. For such a new generation of assessments to market itself in a CALT setting, proponents of CDA are left with the challenging duty of conceptualizing in a clear manner the construct validity, how to measure and evaluate the transfer of skills and the expected long-term positive and even negative washback. Recommended uses of CDA for placement and admission purposes and proposals that learners who learn better with mediation should be placed in higher levels should be supported with sufficient evidence of the validity of inferences and the subsequent actions. Nevertheless, the sound theoretical assumptions on which CDA is built, the types of abilities that can only be assessed within such type of assessment, the notion of validity that takes fairness and equality as major criteria of tests' validation and, above all, the integration of assessment and learning give SLA researchers a strong impetus to thoroughly investigate the potentials of CDA in second language learning.

References

Aljaafreh, Ali, James P. Lantolf (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal* 78: 465–483.

Chapelle, Carol A. (2003). *English Language Learning and Technology: Lectures on Applied Linguistics in the Age of Information and Communication Technology*. Philadelphia, PA: John Benjamins.

Chapelle, Carol A., Dan Douglas (2006). *Assessing Language Through Computer Technology*. Cambridge: Cambridge University Press.

Birjandi, Parviz, Saman Ebadi (2012). Microgenesis in dynamic assessment of L2 learners' sociocognitive development via web 2.0. *Procedia - Social and Behavioral Sciences* 32: 34–39.

Feuerstein, Reuven, Ya'acov Rand, John E. Rynders (1988). *Don't Accept Me As I Am: Helping "Retarded" People to Excel*. New York: Plenum.

Frederiksen, John R., Alan Collins (1989). A systems approach to educational testing. *Educational Researcher* 18(9): 27-32.

Jacobs, Ellen L. (2001). The effects of adding dynamic assessment components to a computerized preschool language screening test. *Communication Disorders Quarterly* 22: 217–226.

Lantolf, James P., Matthew E. Poehner (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics* 1: 49–74.

Lantolf, James P., Steve L. Thorne (2006). *Sociocultural Theory and the Genesis of Second Language Development*. Oxford: Oxford University Press.

Lin, Zheng (2010). Interactive dynamic assessment with children learning EFL in kindergarten. *Early Childhood Education Journal* 37: 279–287.

Loevinger, Jane (1957). Objective tests as instruments of psychological theory. *Psychological Reports* 3: 635-694 (Monograph supplement 9).

McNamara, Tim (1997). "Interaction" in second language performance assessment. Whose performance? *Applied Linguistics* 18, 446-466.

McNamara, Tim (2001). Language assessment as social practice: Challenges for research. *Language Testing* 18(4): 333- 349.

Messick, Samuel (1980). Test validity and ethics of assessment. *American Psychologist*, 35.11: 1012-1027.

Messick, Samuel (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher* 18(2): 5-11.

Messick, Samuel (1990). Validity of test interpretation and use. Princeton, N.J: Educational Testing Service. ETS-RR-90-11

Messick, Samuel (1996). Validity and washback in language testing. *Language Testing* 13: 241-256.

Moss, Pamela A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and practice* 22.4: 13-25.

Poehner, Matthew E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *The Modern Language Journal* 91: 323–340.

Poehner, Matthew E. (2008). *Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting Second Language Development*. Berlin: Springer Publishing.

Poehner, Matthew E. (2011). Validity and interaction in the ZPD: Interpreting learner development through L2 dynamic assessment. *International Journal of Applied Linguistics* 21(2): 244-263.

Poehner, Matthew E., James P. Lantolf (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (CDA). *Language Teaching Research,* 17(3): 323–342.

Porte, Graeme Keith (2010). *Appraising Research in Second Language Learning: A Practical Approach to Critical Analysis of Quantitative Research (2nd ed.)*. Amsterdam: John Benjamins.

Ratner, Carl (1997). *Cultural Psychology: Theory and Methods*. New York: Plenum.

Shabani, Karim (2012). Dynamic assessment of L2 learners' reading comprehension processes: A Vygotskian perspective. *Procedia - Social and Behavioral Sciences* 32: 321-328.

Shrestha, Prithvi, Caroline Coffin (2012). Dynamic assessment, tutor mediation and academic writing development. *Assessing Writing* 17: 55–70.

Sternberg, Robert J., Elena L. Grigorenko (2002). *Dynamic Testing: The Nature and Measurement of Learning Potential.* Cambridge, UK: Cambridge University Press.

Teo, Adeline, Fu Jen (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning and Technology*, 16.3: 10-20.

Vygotsky, Lev S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Author's address:

Iowa State University, Ames, US, 50011
Ross Hall Building
e-mail: zalonazi@iastate.edu

Received: January 22, 2018
Accepted for publication: February 27, 2018