Review article

Developing and using hints in computerized dynamic assessment of a TOEFL *iBT* reading exam

Sahbi Hidri [a,*], Leila Fekri Pileh Roud [b]

[a] Department of English, Faculty of Human & Social Sciences of Tunis, Tunisia
[b] University of Tehran, Kish International Campus, Iran

ARTICLE INFO

ABSTRACT

This study investigated the impacts of computerized dynamic assessment (C-DA) on a TOEFL *iBT* reading exam among 185 upper-intermediate EFL Iranian learners. The exam included five question types: Vocabulary, detail, negative fact, purpose, and inference items whose answers yielded three types of scores: Actual, mediated, and learning potential. Results indicated statistically significant differences between actual and mediated scores with various reading ability levels in using hints in the question types. Even though C-DA improved the scores in the mediated test items and resulted in significant correlations, there was no empirical evidence that C-DA was conducive to a comprehensive diagnosis of the ability in the Zone of Actual Development (ZAD). The study had direct pedagogical and methodological implications by suggesting more individualized and, accordingly, more effective mediation to learners, such as the interactionist approach.

1. Introduction

Static and traditional forms of assessment have been limited in putting together a comprehensive view on the learners' ability where the target of assessment relates to skills and knowledge only. The scope of assessment is narrow in assessing the learning potential (LP) or the assessment context that needs some particular techniques and strategies. Many testing experts (e.g., Alderson et al., 1995) have long hailed standardized diagnosis tests as valid, reliable, and practical despite their limited scope in uncovering the latent cognitive and metacognitive skills and sub-skills the way dynamic assessment (DA) does. Alternatively, anchoring learning in a socio-cultural context is another approach that calls for participants to be part of a learning activity where much scaffolding is needed to diagnose the LP of these learners. Leaners can show evidence of successful learning in the presence of a more competent peer who can engage them in mediation activities and tasks using specific strategies and skills (Vygotsky, 1986), and they have to self-regulate their learning to move from their actual to proximal zone of development (Lantolf and Poehner, 2004). The Zone of Actual Development (ZAD) serves as a diagnosis of what learners currently need to move to the Zone of Proximal Development (ZPD) (Bakhoda and Shabani, 2019a, 2019b; Hidri, 2014; Poehner and Lantolf, 2005). Unlike the ZPD, the ZAD cannot inform us much about the learners' ability (Hidri, 2019).

The Socio-Cultural Theory of mind (SCT) (Vygotsky, 1986), which overtly foregrounds the process in lieu of the product, has showed its merit that language learning is both social and cultural, rather than individual, and that for learning to take place effectively, it has to be undertaken by a mediator or a more competent peer in a socially and culturally contextualized setting (Poehner and Lantolf, 2005). Advocates of DA (e.g., Ableeva, 2008; Anton, 2009; Grigorenko, 2002; Haywood, 2012) have claimed that DA is not essentially meant to help students to obtain better scores; rather, it is aimed at diagnosing the ZAD and ZPD. However, there should be some caution against the ways actual scores are perceived on the assumption that such scores do not inform much about the amount of mediation that learners need. In order to have a more comprehensive diagnosis of the learners' ability, the mediated score that learners get should signal the amount of mediation that they undergo and respond to in the ZPD (Poehner et al., 2015; Vygotsky, 1986). The diagnosis of developmental processes (Poehner and van Compernolle, 2020) is thought to predict individuals' future performance where instruction and assessment are seamlessly merged into one unified activity to the extent that such diagnosis helps ultimately perform solo. Learners sometimes do not profit much from mediation and, therefore, it cannot be conducive to much more learning, especially when their ZAD is well developed and the ability is already matured. When learners with an actual low ability fail to make great leaps in getting higher scores in the

---

mediated performance, this puts into question the efficiency of using DA (Hidri, 2019).

2. Review of the literature

Reading is a principally significant axis for attaining and learning new information (Brown, 1994; Carrell, 1987). It is defined as the learners' ability to understand, interpret, and finally elaborate text information where they interact vigorously with the text, and typically attempt to convey meaning, and use system, linguistic or/and schematic knowledge (Alderson, 2000). Reading is also multifaceted and cognitively challenging, mainly because learners are required to ultimately show attention, perception, memory, and comprehension (Sellers, 2000). It is presumably highlighted that, in contrast to L1, L2 and/or EFL readers read in a way that enhances the reading processes (Kao, 2020a, 2020b). For comprehension to take place, many factors, such as cultural background, students' motivation, and language ability (Sellers, 2000; Vakilifard et al., 2020), should come into play. For L2 cognitive approach, reading is predominantly pre-determined and recognized as a congruently cognitive, intrapersonal problem-solving process synchronized within the readers' prior knowledge (Alderson, 2000). However, views of language learning in this approach have been criticized because the social context of learning is overlooked. The constructivist approach purports that meaning does not lie within the text *per se*; rather it is shaped by the reader's actual and background knowledge (Barbehön, 2020).

Building on the premise that SCT is not individualistic, but rather collective, it is essential then that learners become co-constructors of meaning where cognitive development can come into play only through social interactions with other more able and knowledgeable partners (Aljaafreh and Lantolf, 1994; Lantolf and Thorne, 2006; Poehner et al., 2019a, b). Williams and Burden (1999); Zuengler and Miller (2006). Although language assessment practitioners proved the positive effects of DA in L2 and/or EFL context (e.g., Ableeva, 2008; Anton, 2009; Hidri, 2014, 2017; Poehner and Lantolf, 2005), some nevertheless failed to argue its reliability, validity from a psychometric perspective (Hidri, 2019).

In approaching reading from a DA perspective, Lantolf and Poehner (2004, p. 50) state that DA is "a procedure for simultaneously assessing and promoting development that takes account of the individual's [ZPD]" where mediators should capture the readers' performance with assistance. To this end, intervention calculates the amount of helpful hints required for readers to arrive at a pre-specified endpoint efficiently. One way of doing this is through the use of computers (Crook, 1991). According to Poehner et al. (2015), computerized intervention hints are gradually constructed from implicit to explicit and they serve as gauging readers' ability to respond correctly to current implicit and future explicit mediations (Cho et al., 2020). C-DA uses pre-determined hints to unveil the in-depth learners' ability (Bakhoda and Shabani, 2019a, 2019b; Tzuriel and Shamir, 2002). Poehner (2008) highlights this trend when he states that

> [C-DA] has some advantages, which are not possibly attainable through non-computerized forms of DA: (a) it can be concurrently administered to a large group of students, (b) individuals might be reevaluated as frequently as desired; and finally (c) report of each student's performance is spontaneously produced. (p. 177)

In C-DA, there are some key concepts that need to be defined. First, according to Poehner and Lantolf (2013), the actual scores are the students' unassisted and independent performance. In the current study, actual scores refer to the learners' attempt to respond to multiple-choice questions without any computerized intervention. Poehner and Lantolf (2010, p. 317) emphasize that "mediation must be aimed at those abilities that are in the process of ripening". The learning potential score (LPS) (Poehner and Lantolf, 2013) takes into account the variance between learners' actual score pertaining to their unassisted performance

and the mediated score, the score they get as a result of mediation, using the following formula: *LPS = (2 * mediated score - actual score)/maximum score* (Kozulin and Grab, 2002). In the present study, the mediated scores were calculated by counting the number of hints that learners obtained through the software and then deducting the number of hints from the total scores of each item.

In the Iranian context, several studies have tackled the possible impacts of computerized mediations on developing the reading skills and sub-skills (e.g., Ajideh and Nourdad, 2012; Estaji and Saeedian, 2020), and other skills such as listening (Delvand and Heidar, 2020). Other studies (Attarzadeh, 2011; Birjandi et al., 2011) praised learners' performance improvements in pre-vs. posttests and that there is a significant relationship between text modes scaffolding and learners' ability levels. Other researchers (e.g., Ghafar and Dehqan, 2013) demonstrated that the sociocultural teaching techniques can positively impact reading by improving the LP. Additionally, Shabani (2012) explored the feasibility of C-DA and concluded that C-DA can distinguish between students with very low proficiency levels and how these students use their underlying cognitive abilities to react to computerized assistance. Also, DA had its positive impact on developing EFL learners' grammar competencies (Estaji and Ameri, 2020). Birjandi and Ebadi (2012) observed EFL learners' development of oral abilities and concluded that the higher the ZPD, the less time students spend on tasks while interacting with the mediator. In another study, Pishghadam and Barabadi (2012) explained that while observing the psychometric testing standards of reliability and validity, C-DA helped obtain in obtaining information about learners' LP. In other contexts, many studies have praised the positive impacts of implementing pre-determined hints in diagnosing the learner ability and in developing learning autonomy in C-DA (e.g., Guthke et al., 1997; Poehner et al., 2015; Teo, 2012; Tzuriel and Shamir, 2002).

There has been some debate as to whether DA is an assessment *per se* or just some form of instruction that facilitates the learning process. Some researchers (e.g., Hidri, 2014; Poehner et al., 2019a, b; Poehner, Qin & Yu, 2019b) have argued about the validity and merit of the score a learner gets knowing that this performance is the result of a mediated activity between the learner and a more competent peer, such as teachers or computers. Yet, interpreting the cases of learners whose actual and independent performance is fully developed has not been the focus of DA since this mode of assessment takes into account the learners with a low language ability and mediate them to develop their language ability (Poehner and Lantolf, 2010). Future performance predictions are formed not only based on individuals' current unattended or independent performance but also on the amount and nature of responsiveness to mediation, and this focus serves as a diagnosis of the learners' ability (Poehner, Qin & Yu, 2019b). Development cannot lead to an endpoint, such as obtaining a high score; rather, it is manifested through transcending the learners' current ability level (Esteve, 2018; Lantolf and Poehner, 2004).

3. Problem and rationale

Most often, teachers lack awareness of the principles underlying conceptions and practices of other alternative forms of assessment (Poehner, 2008, 2009) and they tend to disregard the challenges of developing suitable assessment tools, implementing processes, and inferring outcomes (Torrance and Pryor, 1998). Proponents of DA (e.g., Grigorenko, 2002, 2014) commented that traditional assessments are not accurate in evaluating learners' latent abilities and skills and that one way of enhancing reading skills and strategies could be realized through C-DA which, according to Sadeghi (2007), could remain one of the external factors that influence reading. Traditional ways of teaching are a common practice in Iran where teachers are thought to be the only information provider (Birjandi and Hadidi Tamjid, 2010). Researchers (e.g., Hidri, 2014, 2017; 2019; Lantolf and Poehner, 2004, 2006; Murphy, 2011) supported the view that classical assessments do not lend themselves to addressing learners' potential and developmental levels. Most

DA studies are concerned with the theoretical and pertinent aspects of assessment. However, very few studies have addressed the emerging empirical nature to find practical guidelines for DA applications (Sadeghi, 2007; Sadeghi and Khan, 2011). According to Poehner et al. (2015) and Teo (2012), C-DA, for instance, can overcome the shortcomings and weaknesses of administering DA, including its time-consuming nature. Most L2 DA studies were, in essence, interactionist following a sandwich format, and they explored DA implementation on a limited number of candidates while overlooking the psychometric properties of DA. Addressing the impacts of C-DA on the ability of the Iranian EFL learners in reading has not been given its due momentum. We believe that no study has investigated the use of C-DA to assess and promote EFL learners' performance in a reading TOEFL *iBT* exam at an upper-intermediate level using the interventionist approach. The current study then addressed the following questions:

1. To what extent can C-DA reveal the test-takers' learning potentiality in question types of the TOEFL *iBT* reading exam?
2. Is the relationship between candidates' actual and mediated performance in question types significant?

4. Method

For the sake of this study, one-hundred and eighty-five Iranian EFL upper-intermediate participants (85 females and 100 males) whose age ranged from 20 to 36, and who attended TOEFL preparation courses in different institutes of Iran as part of their MA and/or Ph.D. programs were selected for the study through convenience sampling. They belonged to different disciplines such as engineering, management, nursing, biology, accounting, finance, and business. All the participants gave their unconditional consent[1] to take part in the study after the ethical committee[2] had approved our study. For data collection, five steps were considered: a) test preparation, b) piloting, c) hints development, d) software program preparation and description, and finally, e) test administration. Before administering the test, the required explanations were given in Persian. Students were informed about the study objectives and procedure in using the software program to select their answers.

In phase one, a reading placement test was administered to a pool of 270 participants out of which 185 participants, whose IELTS band ranged from 5 to 6 or TOEFL *iBT* 35–87 (ETS, 2010), were selected to attend the TOEFL course and sit for the C-DA exam. The instrument, a C-DA exam uniquely developed for this study, consisted of two original TOEFL reading passages, each of which was followed by eight multiple-choice questions and prompts. The maximum time allotted for the C-DA exam was 80 min.

In phase two, for practical reasons, the original TOEFL *iBT* reading exam that consisted of three reading passages and 42 questions was compacted to include five question types: Vocabulary, detail, negative fact, purpose, and inference. The prepared test was piloted over a group of 30 learners that had approximately similar characteristics as the target group. Cronbach's alpha value was estimated at $\alpha = .81$, and item analysis results showed that no item was faulty, and, therefore, all items were retained. In phase three, the examinees received two passages to read and 16 questions to answer, all of which were about the five-item types. For each item, three hints, graduated from implicit to explicit, were prepared. For validity purposes, all hints were checked and validated by eight judges who had been in charge of teaching TOEFL courses in Iran. A few hints went through some changes, and the final draft was validated and

made available for the participants. The first most implicit hint advocated that the participants carefully read the paragraph or a few sentences again. The second hint was based on the skills and strategies needed to be mastered to answer questions. The final hint gave the correct answer, accompanied by an explanation. If the test-takers' first attempt was correct, they received an explanation of why this answer was correct. An example of the reading questions and hints is given below. The instructions are illustrated in Figures 1, 2, 3, and 4:

"According to paragraph 4, scholars believe that wild cattle, horses, and mammoths are the animals most frequently portrayed in cave paintings for the following reasons EXCEPT:
a. These animals were difficult to hunt because of their unpredictable behavior
b. People preferred these animals for their meat and skins.
c. The painters admired the beauty of these large animals.
d. People feared these animals because of their size and speed.
Hint One: No, that is not the correct answer. Read the paragraph once more and try again.
Hint Two: No, that is not the correct reply. This is a negative fact question. You should look for the option which is not mentioned in the paragraph or which has a detail that is inaccurately mentioned. Read the paragraph once more and try again
Hint Three: No, (c) is the correct answer. Other alternatives are stated in the paragraph directly or indirectly.
If the correct answer is given in the first place:
Yes, (c) is the correct answer. Other alternatives are stated in the paragraph directly or indirectly".

In phase four, a computer software expert designed a software program to set up pre-determined hints so that the exam items could be tested dynamically. The software program comprised of three main parts: First, on the opening page, the examinees needed to fill out a bio-data form (name, age, gender, etc.). Second, as soon as reading the instructions was done, the examinees could begin the test by clicking on the "*NEXT*" button. The test started with the first reading passage which was given in a full-page format so that the test-takers could scroll up and down to read all text paragraphs (Figures 1, 2, 3, and 4).

At the top right of the screen, there was a clock displaying the time left to finish the exam. After finishing reading the passage by clicking on the "*NEXT*" button, the first question appeared on the screen. On top of each question, the paragraph which the question referred to was also given, and students had to tick the correct option. If the first attempt was correct, the test-takers were provided with an explanation to justify the correct answer just before clicking on the "*NEXT*" button to move to the other prompt. Providing this justification was meant to check whether the test-takers guessed the right answer. If the first response was wrong, the test-takers were given a hint after which they had to try again and select the right answer, based on the first most implicit hint. When their reply was correct, they were provided with a justification of viewing an explanation before moving to the next question. However, if the second reply was wrong, a more explicit hint was proposed. This process continued till either the examinees answered correctly or moved to the next prompt. When the examinees did not select any answer within 3 min, they were automatically moved to the next question.

Finally, in phase five, the test generated five sets of scores on the five-item types as well as the actual, mediated, and the learning potential scores (LPSs) that were immediately available upon finishing the exam. The actual score was about the unassisted and independent performances where the examinees received either a maximum score of 3 if they answered correctly, or 0 if they provided the wrong answer. However, the mediated score was weighted in a way that for each mediating prompt a participant received, one point was deducted. Therefore, for any given item, the examinees' actual score was 0 or 3, but the range of the mediated score could be from 0 to 3, depending upon the number of hints they selected. Finally, to estimate the LP, an LPS was calculated for

---

[1] All the participants gave their approval to be part of this study.
[2] The ethical committee that approved data collection was composed of seven members: Dr Sayyed Mohammad Alavi, Dr Shiva Kaivanpanah, Dr AliAkbar Farahani, Dr Hosein Karami, Dr Asfar Rouhi, Dr Hasan Soleimani, Dr Shiva Kaivanpanah. For further details on this approval, please contact the corresponding author of this article.
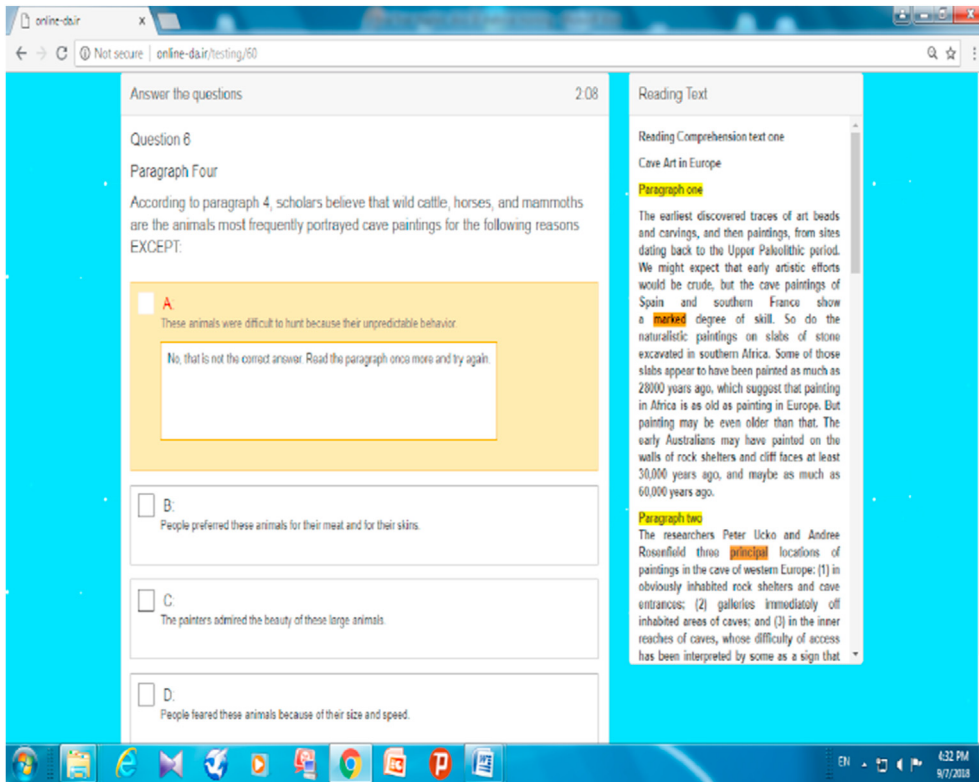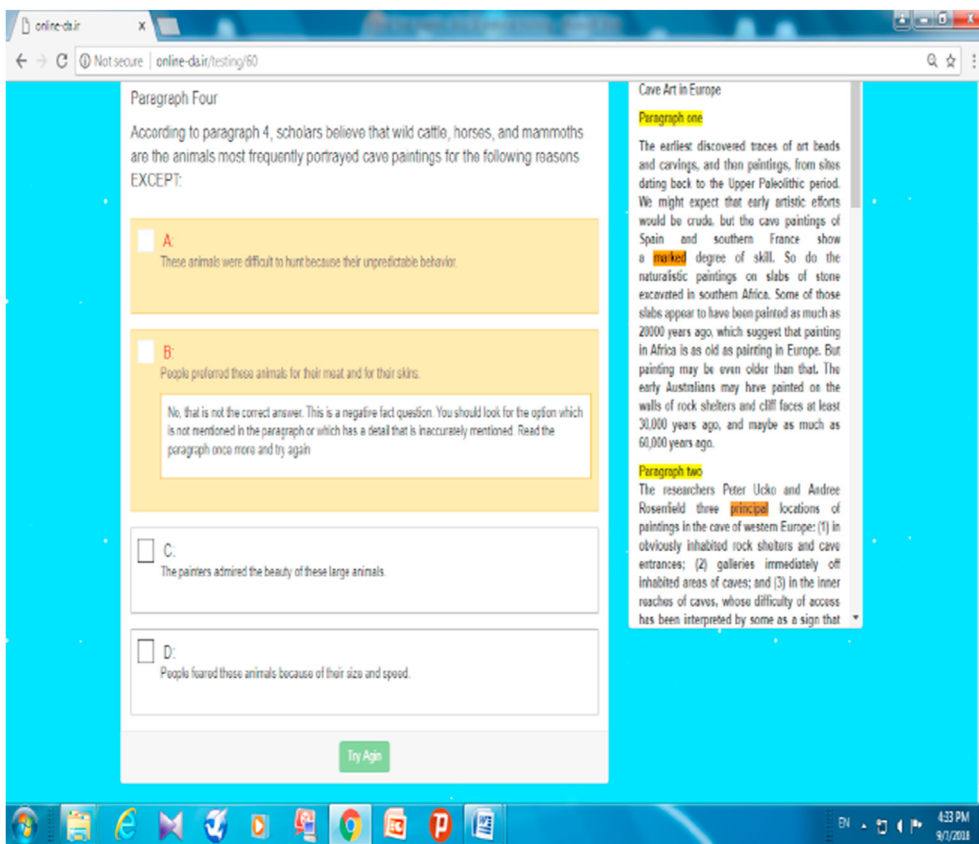
Figure 1. First attempt with the wrong answer.

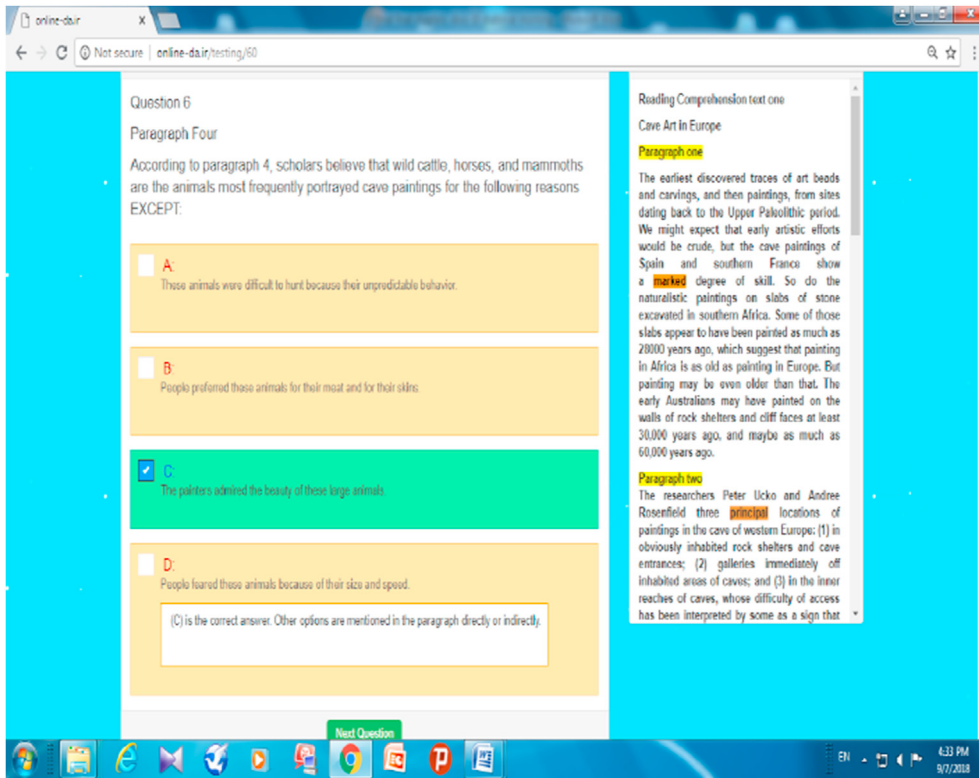**Figure 2.** Second attempt with the wrong answer.

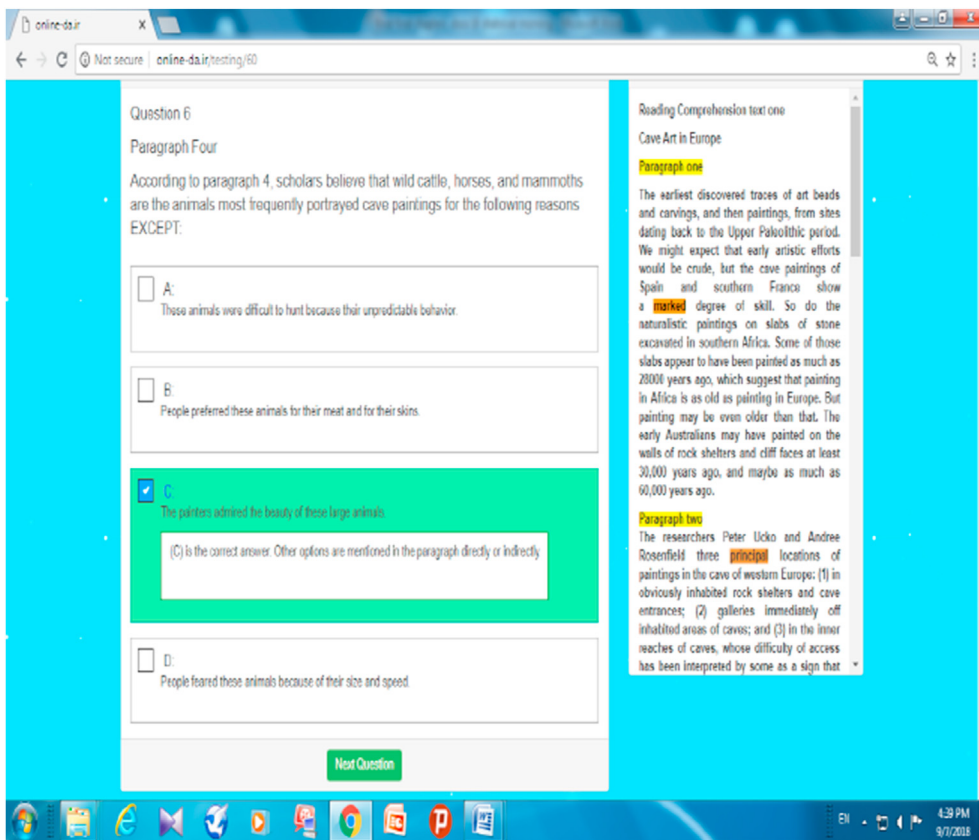**Figure 3.** Third attempt with the wrong answer.

Figure 4. First attempt with the correct answer.

each learner. This score, along with effect size, measured through correlation coefficients, could be a reliable diagnosis of students' reading potential.

The present study adopted a repeated measures design. Before performing the necessary statistical analyses, reliability was estimated through Cronbach's alpha coefficients. To examine the instrument validity, Principal Component Analysis (PCA) was employed, and to compare the actual and mediated performances, a paired sample t-test was performed. However, to form a comprehensive idea about the test-takers LP in question types, each participant's LPS was initially calculated based on the equation by Kozulin and Grab (2002) then two-separate One-Way ANOVA tests were carried out in these question types. Finally, to compare the number of hints in question types across the levels of reading ability, a series of Chi-Square tests were used.

5. Findings

This study examined the C-DA impacts on EFL learners' performance in a TOEFL *iBT* reading exam. The reliability coefficient analysis of all the test-takers' responses to all items was estimated at $\alpha = .798$. The KMO was lower than .6 (.465), and the Bartlett's Test was significant at $p = < .000$. To check the exam construct validity, PCA was performed at the item level for the actual and mediated scores and it showed that the actual and mediated scores of each item loaded onto each factor and that the 16 factors had loadings that ranged from .91 (item 14, actual purpose) to .96 (item 11, actual vocabulary and 15, mediated vocabulary).

Table 1 describes the actual and mediated scores, as well as the results of the paired-sample tests. The actual mean scores of detail, inference, negative facts, and purpose items were relatively the same ranging from 6.00 (detail) to 6.59 (inference), except for vocabulary (8.91). This conveyed that test-takers had more focus on improving their vocabulary knowledge and getting the vocabulary test items correct. The SD actual inference question (9.42) was the highest compared to the other item types, such as detail actual (4.25) and vocabulary (4.18). In the mediated scores, the detail (12.49), negative facts (12.07), and purpose (12.41) means scores were relatively the same. The vocabulary scores (13.27) were the highest, while inference (11.6) was the lowest. The SD inference question (7.28) was higher than those of the other item types. The SD of detail (2.88) and vocabulary (2.97) items were the lowest. This meant that the test-takers' scores in detail and vocabulary items were relatively closer to one another. The gain score (mean mediated score–mean actual score, 1.90–1.08 = .82) showed good progress from the actual to the mediated score.

To find possible differences between candidates' actual and mediated performance, a paired sample t-test was performed, Table 1, where its output shows that the differences between actual and mediated scores were significant ($p = .00$) as per the different question types, including vocabulary (t-value (184) = -22.42, $p = .00$), detail (t-value (184) =

-30.75, $p = .00$), negative fact (t-value (184) = -18.89, $p = .00$), purpose (t-value (184) = -21.07, $p = .00$ and inference (t-value (184) = -11.77, $p = .00$). The total mean score of the t-value of all the item types was estimated at -20.98. The test-takers' mediated scores were more homogeneous. Thus, the null hypothesis that there was no significant relationship between actual and mediated performance in different item types was rejected.

The LPS mean values in various item types, Table 2, indicated that the highest LP was in detail type ($\bar{x} = .64$), followed by purpose ($\bar{x} = .61$), negative fact ($\bar{x} = .56$), inference ($\bar{x} = .50$), and finally, vocabulary ($\bar{x} = .43$) item types. The median scores of detail, negative facts, and purpose items were the same .66 each, and .00 for the inference items, indicating that students needed to have other skills and probably other hints to answer correctly. In addition, the negative inference value (-1.33) of the LPS minimum score implied that the hints in mediated scores distracted the test-takers to select the correct option.

Table 3 describes the relatively low vocabulary LP items (column 1) that ranged from .00 to 1.22, with a mode of .44. This meant that the test-takers tended to employ different traits to answer vocabulary items. The LP of detail items (column 3) ranged from .00 to 1.47, with a mode of .67. Compared to vocabulary, detail items required more condensed traits to be answered correctly. Moreover, 37.9% (19.5 + 18.4, column 4) of test-takers had an LP of .67 and .80, thus, denoting the fact that both the LPS of negative fact and detail items were similar. Moreover, The LP of negative items (column 5) ranged from 0.00 to 1.33 where 38.4% (20.0 + 18.4) (column 6) had an LP of .67 and 1.00 respectively, which denoted that mediated practices were more influential. The purpose items mode (column 7) was higher than detail and negative, with a value of .67, with an LPS that ranged from .33 to 1.33, and it implied that the mediated practices had more influential effects on getting correct answers.

The range of purpose LP was condensed. Also, 42.7% (14.6 + 28.1, column 8) had an LP of .67, implying that mediated practices were also more influential. Interestingly enough, 23.8% (12.4 + 11.4, column 8) had an LP of 1.00 and 1.33 respectively. There were only three levels of inference LP with values ranging from -1.33 to 1.33, and a mode of .67. The negative LP took into consideration the view that one participant had a higher actual than the mediated score, which meant that the test-takers were relatively having similar and limited traits in getting inference items correct. The challenging issue of inference item was mainly related to the high level of LP as 49.7% (23.2 + 26.5) (column 10) of the population possessed LP of the mean scores.

To report the LP differences between the five-item types, the One-Way ANOVA measure, Table 4, was performed. The results postulated a significant difference (f $_{(4,920)}$ = 8.46, $p = < .00$) between the different reading item types, denoting that item types posed different difficulty levels for the test-takers. However, the Scheffe post hoc test of LP scores in the item types (Table 5) did not confirm the significant relationship

Table 1. Mean actual and mediated scores.

	Mean	SD	SEM	t-value	df	sig
Actual detail	6.00	4.25	.31	-30.75	184	.00
Mediated detail	12.49	2.88	.21			
Actual inference	6.59	9.42	.69	-11.77	184	.00
Mediated inference	11.60	7.28	.53			
Actual purpose	6.27	6.31	.46	-21.07	184	.00
Mediated purpose	12.41	4.90	.36			
Actual negative fact	6.37	6.70	.49	-18.89	184	.00
Mediated negative fact	12.07	4.84	.35			
Actual vocabulary	8.91	4.18	.30	-22.42	184	.00
Mediated vocabulary	13.27	2.97	.21			
Mean actual score	1.08	.36	.02			
Mean mediated score	1.90	.28	.02			

Table 2. LPS descriptive statistics in item types (N = 185).

	Vocabulary	Detail	Negative fact	Purpose	Inference
Mean	.43	.64	.56	.61	.50
Median	.44	.66	.66	.66	.00
Mode	.44	.67	.67	.67	.00
SD	.26	.28	.40	.39	.57
Minimum	.00	.00	-1.00	.00	-1.33
Maximum	1.22	1.47	1.33	1.33	1.33

Table 3. LP of *Vocabulary, Detail, Negative Fact, Purpose* and *Reference* items (N = 185).

Vocabulary		Detail		Negative Fact		Purpose		Inference	
LP	%	LP	%	LP	%	LP	%	LP	%
.00	5.9	.00	1.6	.00	18.4	.33	9.2	-1.33	49.7
.11	5.9	.13	1.6	.33	15.1	.33	7.0	.00	23.2
.11	.5	.27	9.2	.33	8.1	.67	14.6	.67	26.5
.11	2.7	.27	2.7	.67	11.9	.67	28.1	1.33	100.0
.22	1.1	.40	14.6	.67	20.0	1.00	12.4	Total	
.22	13.0	.53	3.2	1.00	18.4	1.33	11.4		
.33	8.1	.53	.5	1.33	7.6	Total	100.0		
.33	3.2	.53	5.4	Total	100.0				
.33	5.4	.53	3.2						
.44	16.2	.67	19.5						
.44	2.7	.80	18.4						
.56	7.6	.93	6.5						
.56	3.2	.93	1.6						
.67	5.9	1.07	4.9						
.67	6.5	1.20	6.5						
.78	2.7	1.47	.5						
.78	1.1	Total	100.0						
.89	3.2								
1.00	1.1								
1.00	1.1								
1.11	1.6								
1.22	1.1								
Total	100.0								

Table 4. Comparing LP in item types (N = 185).

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5.49	4	1.37	8.46	.000
Within Groups	149.41	920	.162		
Total	154.91	924			

between LP of vocabulary and inference ($p=<.65$), detail and negative fact ($p=<.46$), details and purpose ($p = .95$), negative fact and purpose ($p=<.88$), negative fact and inference ($p=<.61$), purpose and inference item types ($p=<.12$).

To compare the number of hints across reading ability levels, a series of Chi-Square analyses were conducted (Table 6). Test-takers were classified based on four relative equal groups, while their actual scores were considered as their language ability. Most of the participants (35.67%) clustered around the moderate low reading ability. It could be safe to claim that most test-takers had a low reading ability of almost 63% (27.2 + 35.67). As for the high reading ability, only 17.29 had a high ability. Therefore, the null hypothesis that there was not any statistically significant difference between the number of hints of question types across ability levels could be safely rejected.

6. Discussion

In this study, we explored the C-DA effects on the EFL learners' reading ability. A significant finding of the study was its accentuation of the fact that using pre-determined C-DA to boost examinees' performance was practical and effective, not only in improving their reading performance but also in diagnosing their LP. We stressed the fact that, like other studies, DA of abilities predicted learners' functioning better than when such abilities were measured independently in the ZAD. Assuredly, the ZPD was fixed by the test-takers to take advantage of mediation, where this zone was measured in the context of shared activities. Mediated performances were perceived as key indicators of the status of learners' maturing psychological functions as basic paths, especially when their psychological functions could take advantage of collaborative actions.

Table 5. Scheffe post hoc test of LPS in item types (N = 185).

(a) Fac.Al.Lps	(b) Fac.Al.Lps	Mean Difference (a-b)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Vocabulary	Detail	-.21*	.04	.00	-.34	-.08
	Negative fact	-.13*	.04	.03	-.26	-.00
	Purpose	-.17*	.04	.00	-.30	-.04
	Inference	-.06	.04	.65	-.19	.06
Detail	Negative fact	.07	.04	.46	-.05	.20
	Purpose	.03	.04	.95	-.09	.16
	Inference	.14*	.04	.01	.01	.27
Negative fact	Purpose	-.04	.04	.88	-.17	.08
	Inference	.06	.04	.61	-.06	.19
Purpose	Inference	.11	.04	.12	-.01	.24

* The mean difference is significant at the 0.05 level.

Table 6. Reading ability levels (N = 185).

Level of reading ability	Frequency	Percent
Low	50	27.2
Moderate-low	66	35.67
Moderate-high	37	20
High	32	17.29

Like other studies (e.g., Ahmadi and Barabadi, 2014; Hidri, 2014; Poehner and van Compernolle, 2020), the current findings could be supported by the fact that DA created a supportive atmosphere to highlight candidates' further learning and improvement by considering their ZAD and ZPD.

DA emerged with a solid foundation on SCT. It maintained that having a comprehensive diagnosis of abilities required active intervention on the part of mediators and learners both in the ZAD and ZPD. In this regard, developing the learner ability is contingent not only on the past learning experience, but also on the assessment that the learner is engaged in. This approach shifted its focus from the product of what was already learned to the processes through which abilities could be formed (Poehner, 2009). This study echoed the results of other researchers (e.g., Ahmadi and Hamzavi, 2013; Anton, 2009; Lantolf and Poehner, 2004; Poehner et al., 2015; Poehner, Qin & Yu, 2019b), who praised DA and its diagnostic understanding of students' latent learning problems both through the provision of particular pre-determined hints and prompts and as an assistance to language educators in predicting future performances. This meant that DA could contrive a better understanding of actual reading abilities and how they were developed and used in an online environment. The significant improvement of candidates from actual to mediated performance could be typically attributed to the features of C-DA that precluded the different issues, such as the absence of stimulus and motivation, inhibition failure, and inattention. This result is also echoed in the study conducted by Pishghadam and Barabadi (2012).

As substantiated by Anton (2009), we contended that educators could misrepresent students' abilities if they merely took into account the results of traditional assessments. In this respect, the current findings were in agreement with other works (e.g., Ahmadi and Barabadi, 2014; Haywood and Lidz, 2007; Poehner and Lantolf, 2005; Poehner et al., 2019a, b) that overtly praised C-DA to ultimately reveal test-takers LP. Similarly, the findings presented a significant difference in the candidates' actual and mediated performances in the different question types on the one hand, and between the number of hints in question types and the different reading ability levels, on the other. Significant correlations were also reported by several studies (e.g., Ahmadi and Jafari, 2017; Alavi et al., 2011; Aljaafreh and Lantolf, 1994; Sadeghi and Khan, 2011).

For Poehner and Lantolf (2005), the construct of ZPD implied that potential development differed from the actual one, and this conveyed the veracity of the premise that the latter could not be a predictor of the former. In this study, the developmental LP was not a priori diagnosis. Rather, it emanated from the mediated intervention between students and the computer, and it is in this regard that DA was perceived as adopting a theory-driven approach to assessment since many notions of Vygotsky's theory, specifically ZPD, provided the cornerstone of this approach to assessment. Current results confirmed other studies' results (Estaji and Ameri, 2020; Lidz and Gindis, 2003) in that the construct of cognitive functioning was not fixed but rather a socio-cultural formation hence the SCT dimension of language acquisition and learning. This finding was in agreement with other results (e.g., Esteve, 2018; Grigorenko, 2002; Haywood, 2012; Kao, 2020a, 2020b) that emphasized the positive learning impacts of integrating instruction and assessment where mediators are supposed to implement C-DA to improve language learning. Moreover, several studies (e.g., Nirmalakhandan, 2013; Teo, 2012) proved the positive effects of C-DA in overcoming the weaknesses of DA, including its time-consuming nature. This study seemed to be significant as it may shed light on the efficacy of C-DA in diagnosing and promoting the test-takers' ability in reading and other standardized language skills. According to Poehner and Lantolf (2010), fairness in education necessitated providing appropriate mediation directed to learners' LP in that assessment should be able to assess this size and also describe the learners' ever-changing ability to learn with mediation. This idea was reflected in the results of the current study.

Findings of the study might be interesting to similar-related contexts. However, data of the study showed that some learners did not profit from the pre-determined hints in the same way even though they had similar actual and independent scores. Considering the ZAD, we could not diagnose the amount of mediation that the test-takers need to answer the different item types. In addition, it was difficult to know if the types of mediation students selected were helpful in diagnosing the aspects of their ZAD and ZPD. Mediation should be customized according to the learner ability. This study sparingly used the interventionist approach only, and one of the narrow aspects of this approach, unlike the interactionist one, was its limited scope to reflect the main aspects of SCT where learners and mediators can interactively be engaged in joint

activities to co-construct meaning. The evidence of an interactive dialogue between the learners and mediators might put the results of this study at stake. We observed that carrying out the exam with the help of a computer might not reflect the SCT idea that DA is built on and perhaps the interactionist approach stands in a better position to reflect the core of SCT. One possible overarching rule in any learning environment is that effective learning might not take place by merely exposing learners to online learning materials, and therefore, claiming that C-DA can predict developmental learning should be treated with caution. However, it can still help to form a comprehensive diagnosis of this ability at this level. The absence of careful human use of DA with these learners might inadvertently contaminate the activity and, therefore, all the learning context would fail.

7. Implications and recommendations

Using a software program provided three sets of scores: Actual, mediated and LP, each of which gave some diagnosis of the EFL learners' ability. The actual scores presented students' knowledge of what was already learned, while mediated scores informed us about how much progress happened throughout the test administration, signifying students' responsiveness to the pre-determined hints. Additionally, the LPSs showed how approachable students were to present diagnosis and future instructions. Therefore, conducting classes using a blended form of learning and assessment is one of the pedagogical implications we highlighted. Similarly, one of the methodological implications of the study was the online DA mediation students were engaged in and whose purpose was to diagnose their LP in different question types, hence the necessity to consider similar research instruments in investigating the implementation of DA in an online environment.

Since findings of the present study confirmed that C-DA had significantly positive impacts on improving the test scores, learners could take advantage of such procedure to improve their reading ability, and EFL teachers should raise their students' awareness of the necessity to employ C-DA in learning. ELT practitioners, educationalists, decision makers, and EFL teachers are encouraged to think of different ways to employ C-DA along with traditional and standardized assessments since both assessment types could empower educators with more comprehensive ideas about the curriculum in general.

Careful contemplation of future suggestions could be highlighted. This study tackled the impacts the effects of C-DA on a TOEFL *iBT* reading exam, and it could be replicated using other high-stakes exams among learners with different proficiency levels to check whether the same results would be obtained. By replicating the study using equal numbers of males and females, researchers might preclude gender effects as an intervening variable. The significance of conducting this study lied in encouraging EFL teachers and mediators to use C-DA in such English language classes and other similar-related contexts as potential approaches of addressing the reading problems. There is an amplified need for renovating the educational approach in a vastly technological milieu to equip learners with better reading skills in a restricted period of time. Tracking students' progress using the one-to-one interviews or think-aloud protocols on what was going in the test-takers' minds when selecting hints to receive the outcome of their selection could be another challenging research venue.

8. Conclusion

Advocates of DA have confirmed that in traditional forms of assessment merely those abilities which have already developed are determined, and those abilities that are developing are naturally neglected. Consequently, significant opportunities to mature our assessment are missing. It is noteworthy that dynamic and traditional assessments are not mutually exclusive but rather complementary. Consequently, the effectiveness of C- DA as a supportive assessment tool can be applied to different traditional tests to give a comprehensive picture of language

students' abilities. Teachers should not feel satisfied with focusing on a single actual score. However, despite the vast merits of DA in classrooms, these results cannot be extrapolated to other studies (e.g., Hidri, 2019) who highlighted the limited scope of DA in serving as a diagnosis of the learners' ZAD and ZPD.

Declarations

References

Ableeva, R., 2008. The effects of dynamic assessment on L2 listening comprehension. In: Lantolf, J.P., Poehner, M.E. (Eds.), Sociocultural Theory and the Teaching of Second Languages, pp. 57–86. Equinox.

Ahmadi, A., Barabadi, E., 2014. Examining Iranian EFL learners' knowledge of grammar through a computerized dynamic test. Issues in Language Teaching 3 (2), 161–183.

Ahmadi, S.M., Hamzavi, R., 2013. Dynamic assessment in ELT. Language and Translation Studies 45 (4), 83–103.

Ahmadi, S.M., Jafari, F., 2017. The washback effect of dynamic assessment on grammar learning of Iranian EFL learners. J. Lang. Teach. Learn. 7 (1), 55–68.

Ajideh, P., Nourdad, N., 2012. The immediate and delayed effect of dynamic assessment on EFL reading ability. Engl. Lang. Teach. 5 (12), 141–151.

Alavi, M., Kaivanpanah, S., Shabani, K., 2011. Group dynamic assessment: an inventory of mediational strategies for teaching listening. Journal of Teaching Language Skills 3 (4), 27–58.

Alderson, J.C., 2000. Assessing reading. Cambridge University Press.

Alderson, J.C., Clapham, C., Wall, D., 1995. Language Testing Construction and Evaluation. Cambridge University Press.

Aljaafreh, A., Lantolf, J.P., 1994. Negative feedback as regulation and second language learning in the zone of proximal development. Mod. Lang. J. 78 (4), 465–483.

Anton, M., 2009. Dynamic assessment of advanced second language learners. Foreign Lang. Ann. 42 (3), 576–598.

Attarzadeh, M., 2011. The effect of scaffolding on reading comprehension of various text modes on Iranian EFL learners with different proficiency levels. Social Sciences and Humanities 2 (4), 1–28.

Bakhoda, I., Shabani, K., 2019a. Enhancing L2 learners' ZPD modification through computerized-group dynamic assessment of reading comprehension. Innovat. Lang. Learn. Teach. 13 (1), 31–44.

Bakhoda, I., Shabani, K., 2019b. Bringing L2 learners' learning preferences in the mediating process through computerized dynamic assessment. Comput. Assist. Lang. Learn. 32 (3), 210–236.

Barbehön, M., 2020. Reclaiming constructivism: towards an interpretive reading of the 'social construction framework'. Pol. Sci. 53, 139–160.

Birjandi, P., Ebadi, S., 2012. Micro-genesis in dynamic assessment of L2 learners' socio-cognitive development via web 2.0. Procedia-Social and Behavioral Sciences 32, 34–39.

Birjandi, P., Daftarifard, P., Lange, R., 2011. The effects of dynamic assessment on Rasch item and person hierarchies in second language testing. International Journal of Language Studies 5 (1), 125–140.

Birjandi, P., Hadidi Tamjid, N., 2010. The role of self-assessment in promoting Iranian EFL learners' motivation. Engl. Lang. Teach. 3 (3), 211–220.

Brown, H.D., 1994. Teaching by Principles: an Interactive Approach to Language Pedagog. Prentice-Hall Regent.

Carrell, P.L., 1987. Content and formal schemata in ESL reading. Tesol Q. 21 (3), 461–481.

Cho, E., Compton, D.L., Josol, C.K., 2020. Dynamic assessment as a screening tool for early identification of reading disabilities: a latent change score approach. Read. Writ. 33 (3), 719–739.

Crook, C., 1991. Computers in the zone of proximal development: implications for evaluation. Comput. Educ. 17 (1), 81–91.

Delvand, S.A., Heidar, D.M., 2020. Computerized group dynamic assessment and listening comprehension ability: does self-efficacy matter? Journal of Language and Education 6 (1), 157–172.

Educational Testing Service, 2010. Linking TOEFL iBT scores to IELTS scores—A research report.

Estaji, M., Saeedian, A., 2020. Developing EFL learners' reading comprehension through computerized dynamic assessment. Read. Psychol. 41 (4), 347–368.

Estaji, M., Ameri, A.F., 2020. Dynamic assessment and its impact on pre-intermediate and high-intermediate EFL learners' grammar achievement. Cogent Education 7 (1), 1740040.

Esteve, O., 2018. CBI in teacher education programs in Spain as illustrated by the SCOBA-mediated Barcelona formative model. In: Lantolf, J., Poehner, M.E., Swain, M. (Eds.), The Routledge Handbook of Sociocultural Second Language Development. Routledge, New York, pp. 487–504.

Ghafar, S.R., Dehqan, M., 2013. Sociocultural theory and reading comprehension: the scaffolding of readers in an EFL context. International Journal of Research Studies in Language Learning 2 (3), 67–80.

Grigorenko, E.L., 2002. Dynamic Testing: the Nature and Measurement of Learning Potential. Cambridge University Press.

Grigorenko, E.L., 2014. Dynamic assessment and response to intervention. Two sides of one coin. J. Learn. Disabil. 42 (2), 111–132.

Guthke, J., Beckmann, J.F., Dobat, H., 1997. Dynamic testing: problems, uses, trends and evidence of validity. Educ. Child Psychol. 14, 17–32.

Haywood, H.C., 2012. Dynamic assessment: a history of fundamental ideas. J. Cognit. Educ. Psychol. 11 (3), 217–229.

Haywood, H.C., Lidz, C.S., 2007. Dynamic Assessment in Practice: Clinical and Educational Applications. Cambridge University Press.

Hidri, S., 2014. Developing and evaluating a dynamic assessment of listening comprehension in an EFL context. Lang. Test. Asia 4 (4), 1–19.

Hidri, S., 2017. Specs validation of a dynamic reading comprehension test for EAP learners in an EFL context. In: Hidri, S., Coombe, C. (Eds.), Evaluation in Foreign Language Education in the Middle East and North Africa. Springer International Publishing, pp. 315–337.

Hidri, S., 2019. Static vs. dynamic assessment of students' writing exams: a comparison of two assessment modes. Int. Multidiscip. Res. J. 13 (4), 239–256.

Kao, Y.T., 2020a. A Comparison Study of Dynamic Assessment and Nondynamic Assessment on EFL Chinese Learners' Speaking Performance: Transfer of Learning. English Teaching & Learning, pp. 1–21.

Kao, Y.T., 2020b. Effects of group dynamic assessment on L2 Chinese learners' literacy development: learners' responsiveness to interactive mediation. Appl. Ling. Rev. 1 (ahead-of-print).

Kozulin, A., Garb, E., 2002. Dynamic assessment of EFL text comprehension of at-risk students. Sch. Psychol. Int. 23, 112–127.

Lantolf, J.P., Poehner, M.E., 2004. Dynamic assessment: bringing the past into the future. J. Appl. Ling. 1 (1), 49–72.

Lantolf, J.P., Thorne, S.L., 2006. Sociocultural Theory and the Genesis of Second Language Development. Oxford University Press.

Lidz, C.S., Gindis, B., 2003. Dynamic assessment of the evolving cognitive functions in children. In: Kozulin, A., Gindis, B., Ageyev, V.S., Miller, S.M. (Eds.), Vygotsky's Educational Theory in Cultural Context. Cambridge University Press, pp. 99–116.

Murphy, R., 2011. Dynamic Assessment, Intelligence and Measurement. Wiley-Blackwell.

Nirmalakhandan, N., 2013. Improving problem-solving skills of undergraduates through computerized dynamic assessment. Procedia- Social and Behavioral Sciences 83, 615–621, 2013.

Pishghadam, R., Barabadi, E., 2012. Constructing and validating computerized dynamic assessment of L2 reading comprehension. Iranian Journal of Applied Linguistics 15 (1), 73–95.

Poehner, M.E., 2008. Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting Second Language Development. Springer Publishing.

Poehner, M.E., 2009. Group dynamic assessment: mediation for the L2 classroom. Tesol Q. 43 (3), 471–491.

Poehner, M.E., Lantolf, J.P., 2005. Dynamic assessment in the language classroom. Lang. Teach. Res. 9 (3), 233–265.

Poehner, M.E., Lantolf, J.P., 2010. Vygotsky's teaching assessment dialectic and L2 education: the case for dynamic assessment. *Mind, Culture, and Activity*. Int. J. 17, 312–330.

Poehner, M.E., van Compernolle, R.A., Esteve, O., Lantolf, J.P., 2019a. A Vygotskian developmental approach to second language education. J. Cognit. Educ. Psychol. 17 (3), 238–259.

Poehner, M.E., Lantolf, J.P., 2013. Bringing the ZPD into the equation: capturing L2 development during C-DA. Lang. Teach. Res. 17 (3), 323–342.

Poehner, M.E., Qin, T., Yu, L., 2019b. Dynamic assessment: Co-constructing the future with English language learners. In: Gao, X. (Ed.), Second Handbook of English Language Teaching. Springer International Handbooks of Education. Springer, Cham, pp. 455–476.

Poehner, M.E., van Compernolle, R.A., 2020. Reconsidering time and process in L2 dynamic assessment. In: Poehner, M., Inbar-Lourie, O. (Eds.), Toward a Reconceptualization of Second Language Classroom Assessment. Springer, Cham, pp. 173–195.

Poehner, M.E., Zhang, J., Lu, X., 2015. Computerized dynamic assessment (C-DA): diagnosing L2 development according to learner responsiveness to mediation. Lang. Test. 32 (3), 337–357.

Sadeghi, K., Khan, A.F., 2011. Dynamic assessment of L2 grammar of Iranian EFL learners: the role of mediated learning experience. Int. J. Acad. Res. 3 (2), 931–936.

Sadeghi, K., 2007. The key for successful reader-writer interaction: factors affecting reading comprehension in L2 revisited. Asian EFL J. 9 (3), 198–220.

Sellers, V.D., 2000. Anxiety and reading comprehension in Spanish as a foreign language. Foreign Lang. Ann. 33 (5), 512–521.

Shabani, k., 2012. Computerized dynamic assessment: an interventionist procedure to assess L2 reading. In: National and 3rd International Conference of E-Learning and E-Teaching, 1, pp. 15–24.

Teo, A., 2012. Promoting EFL students' inferential reading skills through Computerized dynamic assessment. Lang. Learn. Technol. 16 (3), 10–20.

Torrance, H., Pryor, J., 1998. Investigating Formative Assessment: Teaching, Learning, and Assessment in the Classroom. Open University Press.

Tzuriel, D., Shamir, A., 2002. The effects of mediation in computer-assisted dynamic Assessment. J. Comput. Assist. Learn. 18 (1), 21–32.

Vakilifard, A., Ebadi, S., Zamani, M., Sadeghi, B., 2020. Investigating demotivating factors in foreign language learners: the case of non-Iranian Persian language learners. Cogent Education 7 (1), 1–14.

Vygotsky, L.S., 1986. Thought and Language. MIT Press.

Williams, M., Burden, R., 1999. Students developing conceptions of themselves as language learners. Mod. Lang. J. 83 (2), 193–201.

Zuengler, J., Miller, E.R., 2006. Cognitive and sociocultural perspectives: two parallel SLA worlds. Tesol Q. 40 (1), 35–58.