Education and Information Technologies (2021) 26:4943–4968
https://doi.org/10.1007/s10639-021-10498-z

Toward a sociocultural approach to computerized dynamic assessment of the TOEFL iBT listening comprehension test

Leila Fekri Pileh Roud[1] · **Sahbi Hidri**[2]

Received: 28 July 2020 / Accepted: 9 March 2021 / Published online: 29 March 2021

Abstract

The current study addressed the impact of computerized dynamic assessment (C-DA) on the TOEFL *i*BT listening comprehension test administered to Iranian EFL learners ($n=185$) who took part in preparation courses on the TOEFL exam in some language centres in Iran. To mediate the test-takers with hints to process the listening questions, a computer software program was developed, and it was meant to produce the following: Actual, mediated, and learning potential scores. Findings of the study indicated that the actual and mediated scores led to significant different mean scores in various listening ability levels in almost all question types. Generally, results highlighted the significant positive impact of C-DA on improving EFL test-takers' performances in the monologue and dialogue tasks. Teachers were recommended to implement C-DA since the information gained from this sociocultural assessment mode empowers them to provide learners with more individualized and accordingly more effective teaching and assessment strategies.

Keywords Computerized dynamic assessment · Listening · Covid-19 · Learning potential · Mediated · Actual scores · Monologue and dialogue tasks · Sociocultural theory of mind

This study was conducted on the Iranian context. It is a good case that supports the contribution of computerized dynamic assessment to language learning and assessment. We hope the reader will find this contribution interesting since it continues hailing the positive aspects of dynamic assessment that advocates of language assessment have been defending.

✉ Sahbi Hidri
sahbihidri@gmail.com

Leila Fekri Pileh Roud
leilafekri@gmail.com

[1] University of Tehran, Tehran, Iran

[2] Department of English- Faculty of Human & Social Sciences of Tunis- University of Tunis, Tunis, Tunisia

1 Introduction

This study put forward an alternative assessment approach and its relationships with learning as advocated in Vygotsky' Socio-Cultural Theory (SCT) of mind (1978), as well as the concepts of carrying out dynamic assessment (DA) using a computer in a TOEFL *i*BT listening test. Dynamic assessment (DA) calls into question the classical perceptions about assessment and teaching by merging instruction and assessment into a seamless task where different manifestations of support and mediation are undertaken to unveil the cognitive and metacognitive strategies of the test-takers (Lidz & Gindis, 2003). The other challenges affecting the association between teaching and assessment lies within the mediators' lack of familiarity with DA conceptions and a myriad of theories underpinning the ways how DA and C-DA are carried out (Poehner, 2008). Therefore, as advocated by Torrance and Pryor (1998), teachers most often face a plethora of selecting and developing the right assessment tools, implementing the relevant processes, and inferring the valid outcomes. Instead, teachers employ several practices and tasks such as cloze tests, group assignments, and tests but with a speculative understanding of how such assessment methods are implemented to inform about the test-takers' ability. Such traditional approaches to assessment are more likely to yield wrong inferences about the test-takers ability, especially when they emanate from a fuzzy overview of the construct, such as listening comprehension.

Although the washback effect explores the influence of assessment on teaching, DA practitioners reverse this association by allocating much more importance to teaching. That is, to connect assessment to teaching, the assessment processes should emerge from a well-established inquiry of the instructional activities and educational performances as carried out in the classroom (Poehner, 2008) to enable teachers to be engaged in a more active role in ascertaining the relevant assessment practices to the learners' potential. Teachers should not limit their testing methods, nor should they test the test-takers' performance with a single final achievement test. Recently, there has been a lot of research on online assessment and how it is used to promote and facilitate learning. For instance, (Wang, 2008) recorded the online leading role learners can have without the mediators' presence. From a Vygotskyan perspective, this sociocultural and active e-learning context is given its due momentum where both learning and assessment are seamlessly merged, using software to put forward hints based on the learners' answers. Vygotsky's theory promotes the process rather the product to understand learning and development where, as Lantolf and Thorne (2006) put it "the potential development varies independently of actual development, meaning that the latter, in and of itself, cannot be used to predict the former" (p. 328).

2 Review of the literature

In language learning, DA applies Vygotsky SCT of mind to assessment, offering new language classroom insights, as well as information to improve interventions. For Grigorenko (2002), DA is not only making a change in the assessment steps but also a shift towards a novel assessment philosophy that focuses on the role of intervention in helping individuals to develop. Lidz (1995) praised the DA test-intervention-retest format to alter the test-takers' learning behaviour. For Poehner (2007), DA is an ongoing and contextualized activity that engages the learners to unveil their underlying potential to make changes to their learning behavior. Vygotsky SCT of mind situates learning within a socio-cultural context, and that human intelligence generates in this context where joint interactions become fundamental in cognitive development. He believed everything is learned by interacting with others and integrating into the individuals' mental structure. The size of the Zone of Proximal Development (ZPD) is determined as per the ability of learners to benefit from support and mediation. The collaborative assessment of the learners' abilities is a predictor of their functioning rather than that type of assessment that measures performance independently. Therefore, the ZPD is assessed within this shared activity. The assisted performance represents the learners' maturity of psychological functions where mediation applies traditional artefacts, notions, and accomplishments (Lantolf & Thorne, 2006) to regulate the individuals' mental and social activity. However, when the learning experience is not mediated, learners might face tremendous difficulties in coming to grips with the learning reality (Feuerstein & Feuerstein, 1991).

The C-DA form used in this study was conducted through a computer where learners received online mediation through sets of prefabricated hints. According to Poehner and Lantolf (2013), the learning potential score (LPS) considers the difference between the learners' actual and mediated scores, using this equation: "*LPS=(2 * mediated score—actual score) / maximum score*" (Kouzlin & Grab, 2002). A mediated score is the culmination of the score emanating from mediation, be it a software program or human mediator (Poehner & Lantolf, 2013). In the current study, the mediated score is computed by subtracting frequencies of hints from the total score of each item. For example, if the total score of the item was three (3), and the learner used two hints to get the correct answer, the mediated score would be one (1). The hints and prompts were gradually organized from implicit to explicit to make the learning ability as malleable as possible. Similar to DA, C-DA is grounded in the theoretical framework of Vygotsky (1978, 1986). With the development of various technological tools, Dixon-Krauss (1996) recommended their application to apply Vygotsky's notions of carrying out lessons that simplify teaching. In interventionist DA, "a prefabricated and fixed set of clues and hints is determined in advance and offered to learners as they move through a test item by item" (Poehner & Lantolf, 2010, p. 318). C-DA provides more in-depth discovery of the learner capabilities (Tzuriel & Shamir, 2002), and it can additionally act as a classroom teacher to mediate the learners in their ZPD (Crook, 1991).

3 Defining listening

Listening is extensively used in second language learning (Scarcella & Oxford, 1992), yet it is the least explicit skill (Vandergrift, 2004). It holds a key role in communication (Mendelsohn, 1994) as "an active and conscious process where the listener constructs meaning by using cues from contextual information and from existing knowledge while relying upon multiple strategic resources to fulfil the task requirement" (O'Malley et al., 1989, p. 19). Interpreting the listening input is contingent upon "the cognitive environment of the listener" (Buck, 2001, p. 29). This makes of listening an active and inferential process (Buck, 2001; Rost, 2002), and for the listening message to be decoded, it has to rely on the prior and linguistic knowledge (Underwood, 1989; Vandergrift, 2007). As stated by Vandergrift (2007), listening is at the heart of learning any language, and it is a problematical active process for the test-taker to distinguish sounds, understand stress and intonation, recognize vocabulary and grammatical structure, and relate them to a particular context. Two cognitive processes are involved in listening: Top-down and bottom-up. While the former is about the interpretation of meaning by means of background knowledge or schemata, the latter linearly involves generating meaning from the smallest to the largest spoken unit of the language (Nunan, 1998). In the current study, listening is operationally defined as the test-takers' ability to get the correct answer using the predetermined hints in an online environment and that this ability is only defined by the mediated score.

The impact of DA on language skills, such as listening, has been investigated throughout the last two decades. For instance, Ableeva (2008) implemented DA and concluded that the difficulties in learning French uncovered the learners' unique ZPD, unlike the case with the non-dynamic pre-test. Likewise, Alavi et al. (2012) contended that group DA prepared for collaboration and interaction by highlighting common practice. Additionally, Hidri (2014) developed a DA listening test, using static and DA approaches, and concluded that DA was conducive to understanding the cognitive and metacognitive listening processes. Ghahremani (2013) tackled the effect of dynamic DA, summative, and formative assessment on learning listening and recorded that learners in the dynamic group outperformed other groups who took part in the study. In addition, Emadi and Arabmofrad (2015) showcased a comprehensive account of interactive listening and found that DA instructions boosted the test-takers to initiate developmental changes. Hashemi Shahraki et al. (2015) carried out a study to estimate test-takers' listening conversational implicatures of pragmatic knowledge and conveyed that the mediational support improved their pragmatic grasp of conversational implicatures. For Wang (2015), DA of listening enhances the amalgamation of assessment and instruction.

Several investigations have identified the possible effect of C-DA on the language skills. For example, Jacobs's (1998) Kidtalk software used a sequence of computerized tasks to assess pre-school learners' ability. Additionally, Birjandi and Ebadi (2012) implemented a similar computerized context to gauge the developmental levels of the oral ability and found a significantly high correlation between the more advanced ZPD and the less time learners spent on interacting with the mediator. Hidri and Pileh Roud (2020) praised the merits of C-DA in developing the reading ability and in making the test-takers more independent in their assessment tasks and activities. Also, Poehner et al. (2015) designed L2 listening online multiple-choice (MC) tests to tag each test item along with implicit to explicit graduated prompts and contended that there existed some significant differences between actual and mediated scores. In the same vein, Heidari and Afgari (2015) addressed a web-based investigation on EFL learners' socio-cognitive progress through DA of listening, and stressed the actual learners' ability, as well as the diagnosis and assessment potential of the listening developmental level.

4 Rationale and problem

The only interventionist study to date on C-DA of listening was carried out by Poehner et al. (2015) where they addressed one of the main issues in applied linguistics, whether assessment and instruction affect each other. That is, what relationship can there be between instruction and assessment? Concerns have arisen over "teaching to the test," and the "power" that tests have gained to control instruction and narrow down the curriculum where teaching and assessment are perceived as mutually exclusive (McNamara, 2001; Shohamy, 2001). It can then be deducted that testing and teaching have been treated as two separate specializations, each having its well-recognized professional journals and conferences. The first step is needed to debate how "interventions based upon the results of dynamic testing provide superior gains" (Elliott, 2003, p. 189). The current identified research problem extensively addresses how intervention can ultimately improve test performance. To do this, the different steps of the study showcase the effectiveness of graduated intervention in leading to the expected testing outcomes when test-takers succeeded in performing in a better way. This was reflected in phases of the study and it was considered in the discussion section. Therefore, this study aimed to answer the following questions:

1. Are the test-takers' actual and mediated scores in the TOEFL *i*BT listening test statistically different?
2. Can C-DA unveil the potentiality of learning in question types?
3. Is there any statistically significant difference between the number of hints across the levels of the listening ability levels in the different question types?

5 Method

5.1 Participants and context

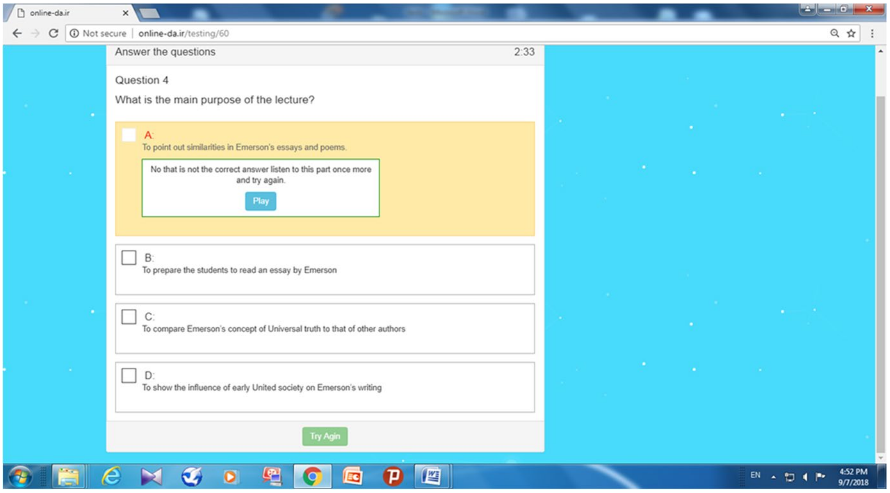
One hundred and eight-five ($n=185$) male and female learners with upper intermediate level in English and who were attending some TOEFL preparation courses in different institutes in Iran, took part in this study. Their age varied from 20 to 36, and they were mainly selected through the convenience sampling procedure. Before being engaged in the current study, a placement test was given to a larger population sample of test-takers only of which 185, who scored IELTS band 5 to 6 or TOEFL *i*BT 87 to 109, were retained to sit for the the TOEFL course. The Listening section duration of the TOEFL *i*BT exam ranges from 60 to 90 min and it includes three conversation passages along with 4 to 6 academic lectures with a length of three to five minutes each. The listening section measures a good variety of skills and sub-skills such as the ability to take notes and infer meaning from academic as well as non-academic contexts. The used instrument was an adapted TOEFL listening test where the test-takers listened to two lectures and dialogues and then received 16 MC questions to answer in 70 min. Data collection consisted of five phases: "*Test preparation, piloting, hints development, software preparation and description, and finally test administration*".

5.2 Phases of the study

The study was carried out in five phases whose purpose was to validate the research instruments and to check the extent to which test-takers' performance changed when it was exposed to a new learning potential. In addition, these phases were meant to check how C-DA provided mediation to the test takers to move from their ZAD and ZPD, while relying on these hints. The different statistical tests used in this study were meant to quantify this progress and to check whether actual and mediated scores as well as hints were significant. We thought that a quantitative analysis of how the test-takers made their progress could be conducive to providing clear evidence on how the relationship between the variables was significant.

In test preparation, the original listening test consisted of three classroom lectures and two dialogues along with 34 test items. For practical reasons, the TOEFL listening test was trimmed. Five question types, main idea, attitude, function, detail, inference, were utilized. The purpose of administering the original TOEFL test here was to expose the test-takers to the five questions and make them familiar with the requirements of the test. In phase two, test piloting, a pool of 30 participants, with similar traits as the target group, sat for the test. Score reliability, Cronbach Alpha, was fixed at 0.823, and the results of item analysis retained all items since no item was found to be faulty. In phase three, hints' preparation, of the dynamic listening test of the test-takers were given two academic lectures and two dialogues and 16 questions to answer. For each of the listening items, three implicit-to-explicit hints were set up by the computer program. For validity considerations, the predetermined hints were first verified, adjusted, and finally validated and approved by some TOEFL teachers ($n = 8$) who were involved in teaching TOEFL preparation courses for many years.

In phase four, software preparation and description, an computer expert designed the software program to test the dynamic listening questions, and this task was carried out by offering the test-takers some prepared hints. The software included three parts: The opening page (Fig. 4), the test (a sample of some answers is presented in Figs. 1, 2 and 3), and the scoring file (Fig. 6). On the opening page (Fig. 4), test-takers were asked to insert their personal details. The next section shortly described the software program (Figs. 5 and 6, see Appendix 1) for a detailed description of the instructions). In section one of the test, the participants listened to a three-minute dialogue, followed by the first question. In case the first trial was correct, an explanation was displayed before proceeding to the next item. If the second trial was not relevant, i.e., incorrect, more explicit details, in the form of a hint, were displayed. However, if the test-takers failed to select one answer within three minutes,



Answer the questions 2:33

Question 4

What is the main purpose of the lecture?

□ A:
To point out similarities in Emerson's essays and poems.

No that is not the correct answer listen to this part once more and try again.

□ B:
To prepare the students to read an essay by Emerson

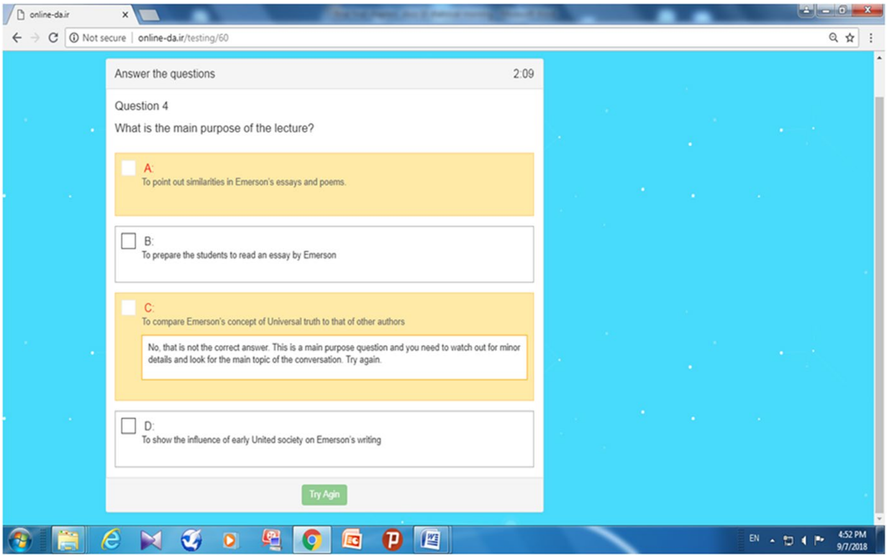
□ C:
To compare Emerson's concept of Universal truth to that of other authors

□ D:
To show the influence of early United society on Emerson's writing

Fig. 1 Wrong answer in the first attempt

they were systematically moved to the next question after the answer was marked as wrong. The following is an example of the hints:

> "What is the main purpose of the lecture?
>
> a. To point out similarities in Emerson's essays and poems.
> b. To prepare the students to read an essay by Emerson



Answer the questions 2:09

Question 4

What is the main purpose of the lecture?

□ A:
To point out similarities in Emerson's essays and poems.

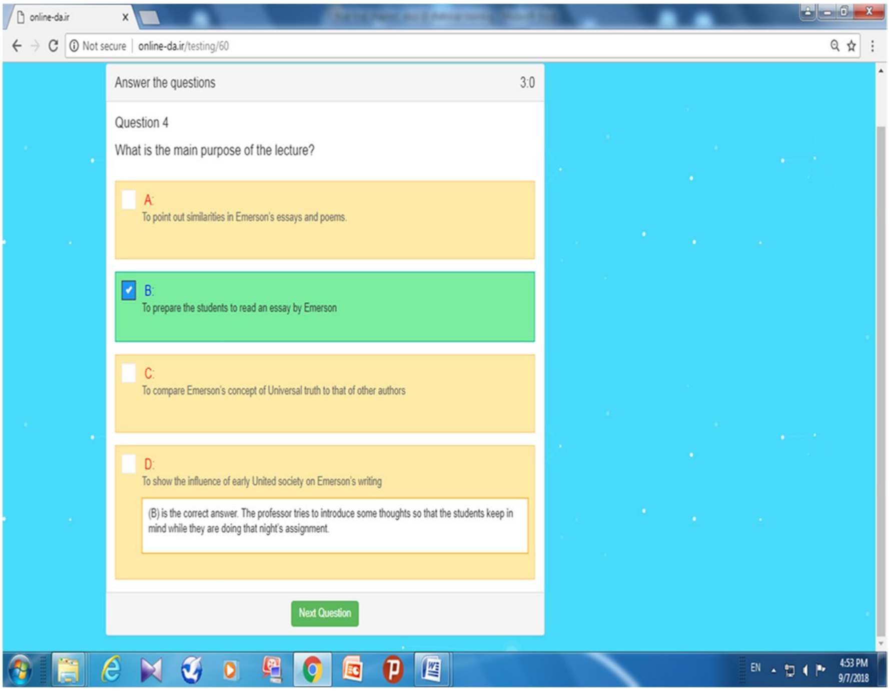
□ B:
To prepare the students to read an essay by Emerson

□ C:
To compare Emerson's concept of Universal truth to that of other authors

No, that is not the correct answer. This is a main purpose question and you need to watch out for minor details and look for the main topic of the conversation. Try again.

□ D:
To show the influence of early United society on Emerson's writing

Fig. 2 Wrong answer in the second attempt



Answer the questions 3:0

Question 4

What is the main purpose of the lecture?

□ A:
To point out similarities in Emerson's essays and poems.

☑ B:
To prepare the students to read an essay by Emerson

□ C:
To compare Emerson's concept of Universal truth to that of other authors

□ D:
To show the influence of early United society on Emerson's writing

(B) is the correct answer. The professor tries to introduce some thoughts so that the students keep in mind while they are doing that night's assignment.

Next Question

Fig. 3 Wrong answer in the third attempt

c. To compare Emerson's concept of Universal truth to that of other authors
d. To show the influence of early united society on Emerson's writing"

"Hint One: No, that is not the correct answer. Listen to this part of the lecture once more and try again."
"Hint Two: No, that is not the correct answer. This is a main purpose question and you need to watch out for minor details and look for the main topic of the conversation. Try again."
"Hint Three: No, (b) is the correct answer. The professor tries to introduce some thoughts so that the students keep in mind while they are doing that night's assignment."
"If the correct answer is chosen in the first place:"
"Yes, (b) is the correct answer. The professor tries to introduce some thoughts so that the students keep in mind while they are doing that night's assignment"

In phase five, scoring procedure, upon finishing the test, the software program yielded the following types of scores: Actual, mediated, and LP. The actual score received three (3) marks if the test-takers answered correctly; if not, it would be zero (0). However, in the mediated score, for any mediating hint a test-taker got, there was a deduction of one mark so that items were graded as follows: actual score

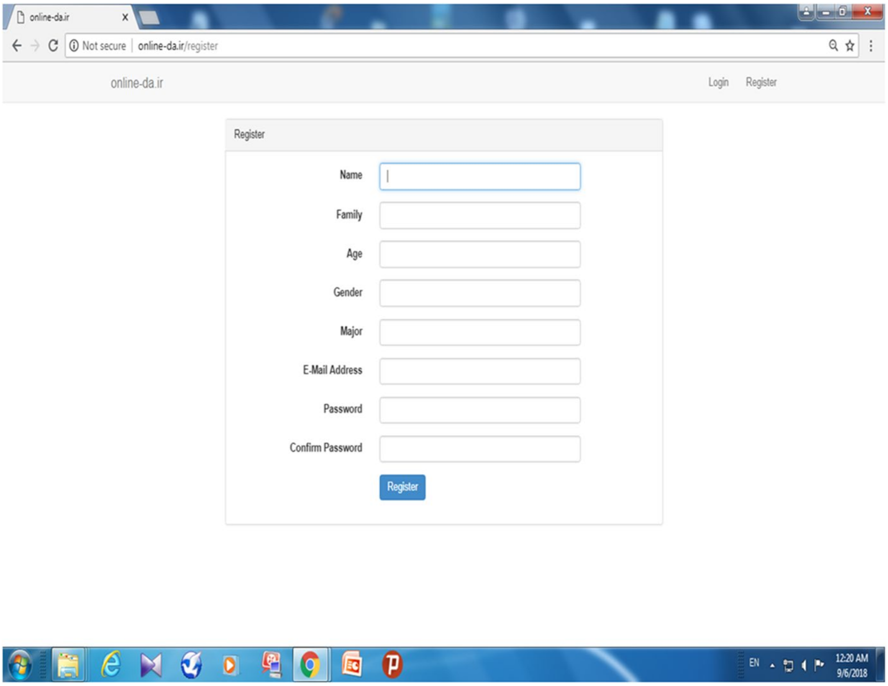
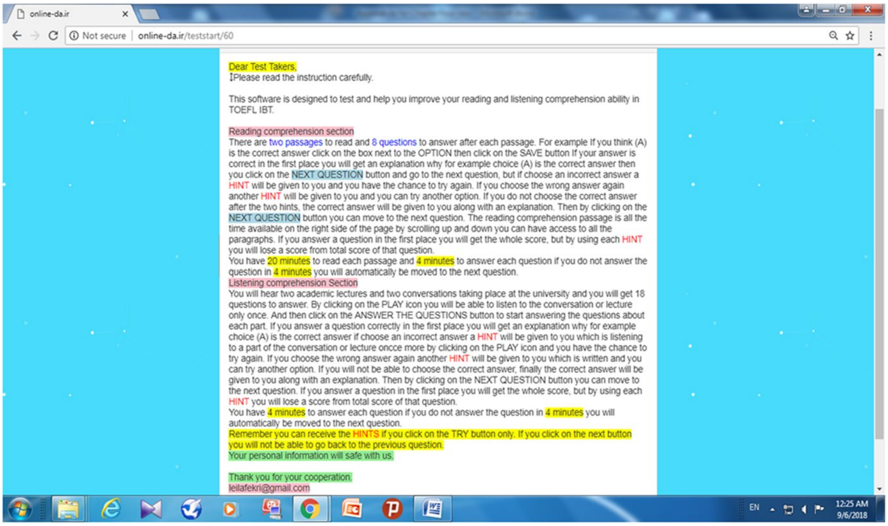


Fig. 4 Opening page



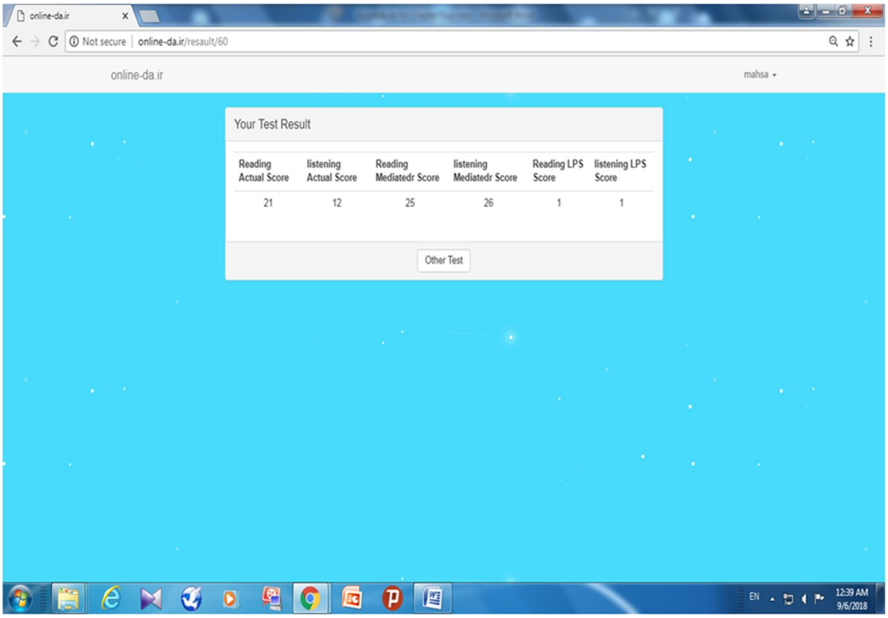
Dear Test Takers,
Please read the instruction carefully.

This software is designed to test and help you improve your reading and listening comprehension ability in TOEFL IBT.

Reading comprehension section
There are two passages to read and 8 questions to answer after each passage. For example If you think (A) is the correct answer click on the box next to the OPTION then click on the SAVE button If your answer is correct in the first place you will get an explanation why for example choice (A) is the correct answer then you click on the NEXT QUESTION button and go to the next question, but if choose an incorrect answer a HINT will be given to you and you have the chance to try again. If you choose the wrong answer again another HINT will be given to you and you can try another option. If you do not choose the correct answer after the two hints, the correct answer will be given to you along with an explanation. Then by clicking on the NEXT QUESTION button you can move to the next question. The reading comprehension passage is all the time available on the right side of the page by scrolling up and down you can have access to all the paragraphs. If you answer a question in the first place you will get the whole score, but by using each HINT you will lose a score from total score of that question.
You have 20 minutes to read each passage and 4 minutes to answer each question if you do not answer the question in 4 minutes you will automatically be moved to the next question.
Listening comprehension Section
You will hear two academic lectures and two conversations taking place at the university and you will get 18 questions to answer. By clicking on the PLAY icon you will be able to listen to the conversation or lecture only once. And then click on the ANSWER THE QUESTIONS button to start answering the questions about each part. If you answer a question correctly in the first place you will get an explanation why for example choice (A) is the correct answer if choose an incorrect answer a HINT will be given to you which is listening to a part of the conversation or lecture oncce more by clicking on the PLAY icon and you have the chance to try again. If you choose the wrong answer again another HINT will be given to you which is written and you can try another option. If you will not be able to choose the correct answer, finally the correct answer will be given to you along with an explanation. Then by clicking on the NEXT QUESTION button you can move to the next question. If you answer a question in the first place you will get the whole score, but by using each HINT you will lose a score from total score of that question.
You have 4 minutes to answer each question if you do not answer the question in 4 minutes you will automatically be moved to the next question.
Remember you can receive the HINTS if you click on the TRY button only. If you click on the next button you will not be able to go back to the previous question.
Your personal information will safe with us.

Thank you for your cooperation.
leilafekri@gmail.com

Fig. 5 Instruction page of the online test



Your Test Result

Reading Actual Score	listening Actual Score	Reading Mediatedr Score	listening Mediatedr Score	Reading LPS Score	listening LPS Score
21	12	25	26	1	1

Fig. 6 Actual and mediated scores

ranged from 0 to 3, but their mediated score could extend from zero (0) to three (3), which was contingent upon the used number of hints. The individual LPS was calculated, using Kozulin and Garb formula (2002). In the last phase, test administration, the test was administered after a formal approval had been received from the ethical committee. Prior to test administration, the test directions and details were explained in Persian to inform the test-takers about nature and purpose of the test and the procedure to the software program to answer all the questions.

6 Results

The reliability coefficient analysis (Table 1) displays a good Cronbach's alpha value ($\alpha = 0.752$) in the two types of scores, actual and mediated, totaling 32. To examine the construct validity of these scores, the PCA of the rotated component matrix of the actual and mediated scores designated that the factors loaded at higher values ranging from 0.90 to 0.95 (see Appendix Table 10 for factor analysis results). Table 2 reveals descriptive statistics of actual and mediated scores. The actual scores had values of 7.35 (SD = 9.07), 7.63 (SD = 4.17), 7.81 (SD = 9.22), 7.44 (SD = 4.73), and 9.02 (SD = 4.77) for attitude, detail, function, inference and main idea,

Table 1 Reliability statistics of actual and mediated scores ($n = 185$)

Cronbach's alpha	Number of items
0.752	32

Table 2 Total actual and mediated scores ($n = 185$)

	Total actual scores					Total mediated scores				
	Min	Max	Mean	SE	SD	Min	Max	Mean	SE	SD
Attitude	0.00	20.00	7.35	0.66	9.07	0.00	20.00	10.41	0.58	7.91
Detail	0.00	17.78	7.63	0.30	4.17	0.00	17.78	10.68	0.33	4.52
Function	0.00	20.00	7.81	0.67	9.22	0.00	20.00	11.74	0.56	7.73
Inference	0.00	18.33	7.44	0.34	4.73	0.00	20.00	10.16	0.37	5.12
Main idea	0.00	20.00	9.02	0.35	4.77	0.00	20.00	10.51	0.40	5.45

respectively. The mean of the mediated scores in attitude (M = 10.41; SD = 7.91), detail (M = 10.68; SD = 4.52), inference (M = 10.16; SD = 5.12), and main idea (M = 10.51; SD = 5.45) items were lower than the function items (M = 11.74; SD = 7.73).

Table 3 describes the paired t-test, and it shows a statistically significant difference in the actual and mediated scores (column 8). The table also shows that the mediated mean scores outperformed the actual ones. The actual means ran from 7.35 (attitude) to 9.02 (main idea), and from 10.16 (inference) to 11.74 (function) for the mediated means. However, the homogeneity of scores in actual and mediated responses of each candidate was more or less the same.

*SEM: Standard error mean.

Figure 7 introduces descriptive statistics of actual and mediated scores of the dialogue and monologue listening input, moving from 6.81 (column 2) and 6.49 to 12.45 and 13.22, respectively.

Concerning the multiple comparisons of the of actual and mediated scores in dialogue and monologue tasks, Table 4, there was no significant difference (column 5, with values of 0.80 and 0.10) between the actual and mediated scores in dialogue

Table 3 Actual and mediated scores ($n = 185$)

		Mean	SD	SEM*	t-value	df	Sig
Pair 1	Main.idea.actual	9.02	4.77	0.35	−3.90	184	0.00
	Main.idea.mediated	10.51	5.45	0.40			
Pair 2	Function.actual	7.81	9.22	0.67	−5.25	184	0.00
	Function.mediated	11.74	7.73	0.56			
Pair 3	Attitude.actual	7.35	9.07	0.66	−4.53	184	0.00
	Attitude.mediated	10.41	7.91	0.58			
Pair 4	Inference.actual	7.44	4.73	0.34	−5.94	184	0.00
	Inference.mediated	10.16	5.12	0.37			
Pair 5	Detail.actual	7.63	4.17	0.30	−6.09	184	0.00
	Detail.mediated	10.68	4.52	0.33			

*SEM: Standard error mean

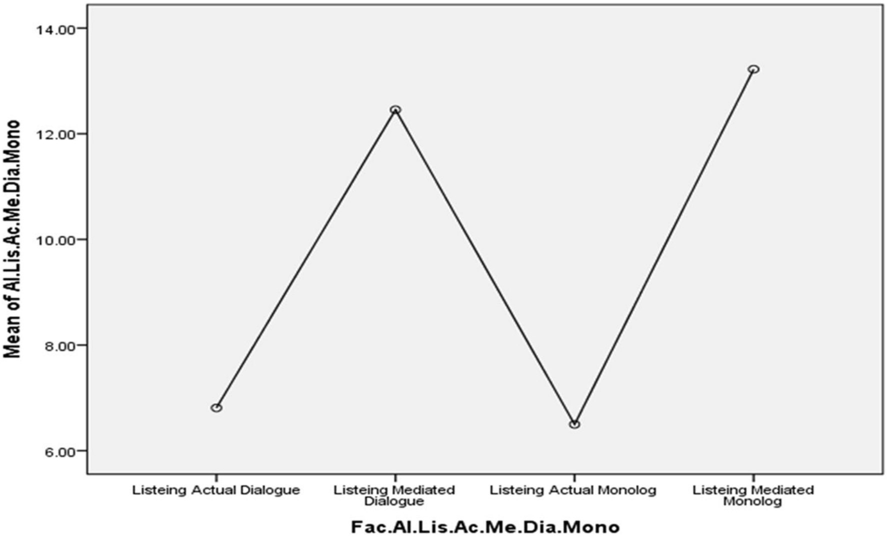


Fig. 7 Dialogue and monologue actual and mediated scores

and monologue contexts, while it was not the case between the other pairs of comparisons with a significant level of *p* = *< 0.000.*

The descriptive statistics of the learning potential scores in the different types of items, Table 5, showed that the learning potential score in function item types ($\bar{x}$ =0.39) was the highest, followed by attitude ($\bar{x}$ =0.30), detail ($\bar{x}$ =0.30), inference ($\bar{x}$ =0.27), and main idea ($\bar{x}$ =0.14) item types. The median scores of detail item types were the highest (0.55), followed by inference and main idea item types (0.23 and 0.22). The median of function and attitude item types was zero (0). The minimum LP of all item types was negative with values of −1.33, −2.00, −1.50 and −1.22 due to other latent variables that distracted the test-takers from answering correctly.

Table 4 Dialogue and monologue comparisons of actual and mediated scores ($n = 185$)

(I) Fac.al.actual mediated.dialogue Monologue	(J) Factor.analysis Actual.mediated. dialogue Monologue	(I-J) Mean Difference	SE	Sig	95% confidence interval	
					Lower bound	Upper bound
Actual dialogue	Mediated dialogue	–5.64*	0.31	0.00	–6.51	–4.76
	Actual monologue	0.312	0.31	0.80	–0.56	1.18
	Mediated monologue	–6.41*	0.31	0.00	–7.28	–5.53
Mediated dialogue	Actual monologue	5.95	0.31	0.00	5.08	6.82
Actual monologue	Mediated monologue	–0.76	0.31	0.10	–1.64	0.10
	Mediated monologue	–6.72*	0.31	0.00	–7.59	–5.85

* A significant difference of the mean is set up at the 0.05 level

Table 5 Descriptive statistics of learning potential scores ($n = 185$)

	Main idea	Function	Attitude	Inference	Detail
Mean	0.14	0.392	0.30	0.27	0.30
Median	0.22	0.00	0.00	0.33	0.55
Mode	0.00	0.00	0.00	0.50	0.89
SD	0.51	1.01	0.919	0.62	0.679
Min	−1.33	−2.00	−2.00	−1.50	−1.22
Max	0.89	2.00	2.00	1.33	1.33

Table 6 gave data of the LP in the five-question types. The LP of the main idea items went from −1.33 to. 89 with 16.2% of the test-takers who had an LP of 0.00. The function LP items varied from −2 to 0.2, and the mode of function question types was 1.33, which represented a high LP and positive effect of mediation. The function LP item type indicated that the mediated practices were influential in reaching the correct answers. The LP of attitude items differed from −2 to 0.2 with a mode of 0.00. The attitude LP specified that 54.1% of the test-takers had an LP below 0.00. The LP of inference items extended from −1.33 to 1.33 with a mode of 0.50. The detail LP items ranged from −1.22 to 1.33 with a mode of 0.89. The purpose of the series of Chi-Square test analyses (Table 7) was to set up a clear comparison in the number of hints across the ability levels. The test-takers were ranked on the basis of four relatively equal groups, ranging from low to high, while the actual scores were examined based on their language ability. The majority of the test-takers clustered around the moderate and moderate high level with a total of 48.64%, and a low percentage of 14.0 of high achievers.

The results of a series of Chi-square analyses of the mediated scores, Table 8, showed that there were statistically significant differences (*p* = < *0.00, p* = < *0.01*) in using hints in the main idea dialogue, function dialogue, attitude dialogue, detail monologue, inference monologue, main idea dialogue, inference dialogue, detail monologue questions types. However, no statistically significant differences were found in detail monologue (19), inference dialogue (24), and main idea monologue (16) question types. A series of pair t-test was carried out, Table 9, and it represented a statistically significant difference in actual and mediated scores so the null hypothesis that the hints could support the test-takers could be safely rejected. The descriptive statistics confirmed that the mean scores in the mediated scores outperformed their actual scores.

The sample comparison, Table 9 (8 pairs out of 16), made clear that the statistically significant difference between actual and mediated scores was linked to the various question types including the main idea dialogue, pair 1, column 6, (T (184) = −13.9, *p* < = *0.00*), function dialogue, pair 2, (T (184) = −13.7, *p* < = *0.00*), attitude monologue, pair 3, (T (184) = −14.8, *p* < = *0.00*), main idea monologue, pair 4, (T (184) = −13.7, *p* < = *0.00*), detail monologue, pair 5, (T (184) = −16.8, *p* < = *0.00*), inference monologue, pair 7, (T (184) = −15.6, *p* < = *0.00*), attitude dialogue (T (184) = −14.8, *p* < = *0.00*), inference dialogue, pair 10, (T (184) = −15.3, *p* < = *0.00*),

Table 6 LP of the questions types ($n = 185$)

Main idea		Function		Attitude		Inference		Detail	
LP	%	LP	%	LP	%	LP	%	LP	%
−1.33	0.5	−2.00	7.6	−2.00	5.4	−1.33	1.1	−1.22	1.1
−1.33	1.1	−1.33	1.1	−1.33	3.8	−1.17	0.5	−1.11	2.2
−1.11	2.7	−0.67	8.1	−0.67	6.5	−1.00	3.8	−1.00	2.2
−0.89	4.3	0.00	34.1	0.00	38.4	−0.83	1.1	−0.89	2.7
−0.67	4.3	0.67	10.3	0.67	17.3	−0.83	1.1	−0.78	1.6
−0.44	0.5	0.67	0.5	1.33	24.9	−0.67	3.8	−0.78	2.2
−0.44	3.8	1.33	31.9	1.33	2.2	−0.67	1.6	−0.67	1.1
−0.22	2.2	1.33	2.2	2.00	1.6	−0.50	3.8	−0.67	2.2
−0.22	1.1	2.00	4.3	Total	100.0	−0.33	1.6	−0.56	2.2
−0.22	0.5	Total	100.0			−0.33	1.1	−0.56	2.2
0.00	16.2					−0.17	2.2	−0.44	3.8
0.22	12.4					−0.17	0.5	−0.33	1.6
0.22	0.5					0.00	8.6	−0.33	0.5
0.22	4.9					0.17	1.6	−0.33	0.5
0.44	3.2					0.17	1.1	−0.22	1.6
0.44	2.2					0.17	3.8	−0.11	0.5
0.44	15.1					0.33	10.8	0.00	0.5
0.44	3.2					0.33	1.6	0.22	1.1
0.67	11.9					0.33	2.2	0.33	4.3
0.67	4.9					0.50	13.0	0.33	0.5
0.89	2.2					0.67	4.9	0.33	1.6
0.89	2.2					0.67	11.4	0.44	7.6
Total	100.0					0.83	3.8	0.44	2.7
						0.83	0.5	0.56	6.5
						1.00	8.6	0.56	3.8
						1.17	1.1	0.67	3.8
						1.33	4.3	0.67	6.5
						Total	100.0	0.78	2.2
								0.78	5.4
								0.89	13.5
								1.00	3.8
								1.11	5.4
								1.22	1.6
								1.33	0.5
								Total	100.0

detail dialogue, pair 12, (T (184) = −12.6, $p < = 0.00$), and purpose monologue (T (184) = −13.5, $p < = 0.00$).

Table 7 Test-takers' listening ability levels ($n = 185$)

Ability level	Frequency	Percent
Low	34	18.37
Moderate low	34	18.37
Moderate high	56	30.27
High	26	14.05
Total	185	100

7 Discussion

Results of the study proved that a full understanding of the listening ability required active intervention in its development and, accordingly, this shift from the classical view of assessment to a more interventionist one, by accentuating the process rather than the product. The current study echoed the findings of other studies on DA (e.g., Poehner et al., 2015) who argued that DA offers a diagnostic understanding of the test-takers' difficulties by providing well-defined and pre-determined hints during the assessment process. The results suggested that C-DA, i.e., the integration of teaching and assessment using a computer program with pre-determined hints, could serve as a mere diagnosis of the test-takers' ability both in the ZAD and ZPD. This integration was efficient in probing the test-takers' LP. Such findings could be justified by the effective implementation of contextualised exam instructions to boost learners' involvement in the task by judging their ZAD and ZPD (Ahmadi & Barabadi, 2014). The test-takers' actual-to-mediated gradual improvement was significant, and it can be essentially attached to the aspects of C-DA that could eliminate any possible learning handicaps. DA procedures were useful in activating the metacognitive listening strategies, and this was in line with other studies (e.g., Ajideh et al., 2012; Alavi et al., 2012; Haywood & Lidz, 2007; Pishghadam & Barabadi, 2012; Poehner, 2007).

Concerning the question types' LP, the findings brought to light the fact that the highest LP was in the function items, followed by the main idea, attitude, detail, and finally inference items. The wide variety of the learning potential scores in the detail part meant that the items were more accessible to the test-takers. However, in other instances, the test-takers lacked other hints to answer correctly. Therefore, gaining information about this potential could allow the language learners to have an accurate picture of their capabilities (e.g., Peña et al., 2001). This aspect was reflected in the current study, especially when the C-DA hints spontaneously promoted the test-takers to self-assess their ability. Also, a statistically significant difference was found between high, moderate-high, moderate-low, and low achievers in the number of hints used in almost all questions' types, except for the inference and detail items. This showed that the test-takers in the diverse listening ability levels tended to have recourse to their multifarious traits to answer the main idea, attitude, and function items, while mainly relying on the allocated hints to answer detail and inference items.

Table 8 A sample of the significance level hints in question types across ability levels ($n = 185$)

	Mediated ability levels	N	Mean rank	Chi-square	df	sig
1.Mainidea.dialogue	Low	34	56.24	15.215	3	0.00
	Moderate low	34	67.74			
	Moderate high	56	85.73			
	High	26	88.81			
2.Function.dialogue	Low	34	49.29	20.705	3	0.00
	Moderate low	34	74.34			
	Moderate high	56	85.99			
	High	26	88.69			
3.Attitude.dialogue	Low	34	58.59	17.690	3	0.00
	Moderate low	34	62.82			
	Moderate high	56	85.07			
	High	26	93.58			
	Total	150				
4.Main.idea.monologue	Low	34	63.53	5.035	3	0.16
	Moderate low	34	81.78			
	Moderate high	56	74.66			
	High	26	84.75			
	Total	150				
5.Detail.monologue	Low	34	59.74	10.598	3	0.01
	Moderate low	34	75.13			
	Moderate high	56	76.90			
	High	26	93.58			
	Total	150				
6.Detail.monologue	Low	34	53.29	21.475	3	0.00
	Moderate low	34	80.56			
	Moderate high	56	74.04			
	High	26	101.08			
	Total	150				
7.Inference.monologue	Low	34	59.15	16.217	3	0.00
	Moderate low	34	64.21			
	Moderate high	56	83.18			
	High	26	95.12			
	Total	150				
8.Main.idea.dialogue	Low	34	59.09	13.805	3	0.00
	Moderate	34	69.29			
	Moderate high	56	79.41			
	High	26	96.65			
	Total	150				

Table 8 (continued)

	Mediated ability levels	N	Mean rank	Chi-square	df	sig
9.Inference.dialogue	Low	34	62.04	5.315	3	0.15
	Moderate low	34	77.54			
	Moderate high	56	78.15			
	High	26	84.71			
	Total	150				
10.Inference.dialogue	Low	34	52.68	17.422	3	0.00
	Moderate low	34	72.62			
	Moderate high	56	82.64			
	High	26	93.73			
	Total	150				
11.Detail.dialogue	Low	34	63.91	4.183	3	0.24
	Moderate	34	80.78			
	Moderate high	56	76.04			
	High	26	82.58			
	Total	150				
12.Detail.monologue	Low	34	59.50	12.170	3	0.00
	Moderate low	34	71.79			
	Moderate high	56	77.86			
	High	26	96.19			
	Total	150				
13.Detail.monologue	Low	34	66.90	7.699	3	0.05
	Moderate low	34	77.49			
	Moderate high	56	70.92			
	High	26	94.02			
	Total	150				
14.Detail.monologue	Low	34	62.34	4.704	3	0.19
	Moderate low	34	80.21			
	Moderate high	56	78.11			
	High	26	80.94			
	Total	150				

The results highlighted the significant and positive difference in the two types of performance, actual and mediated, in the monologue and dialogue parts. However, like other studies (e.g., Ableeva, 2008; Poehner et al., 2015; Shrestha & Coffin, 2012), there was no significant difference in actual and mediated scores in the dialogue and monologue contexts. This implicated that the test-takers brought into play their cognitive abilities in taking advantage of the hints in the monologue and dialogue types. Consequently, assessing listening in the dialogue and monologue contexts dynamically involved the test-takers in joint activities to overcome task difficulty and attain the level where they could construct

Table 9 A sample of mediated and actual scores as per question types ($n = 185$)

Question types		Mean	SD	SEM*	t-value	df	sig
Pair 1	1.Actual.mainidea.dialogue	1.02	1.42	0.10	−13.9	184	0.00
	1.Mediated.mainidea.dialogue	1.99	0.96	0.07			
Pair 2	2.Actual.function.dialogue	1.00	1.41	0.10	−13.7	184	0.00
	2.Mediated.function.dialogue	2.01	0.93	0.06			
Pair 3	3.Actual.attitude.dialogue	1.02	1.42	0.10	−14.8	184	0.00
	3.Mediated.attitude.dialogue	2.05	0.89	0.06			
Pair 4	4.Actual.mainidea.monologue	0.92	1.38	0.10	−13.7	184	0.00
	4.Mediated.mainidea.monologue	1.88	1.00	0.07			
Pair 5	5.Actual.detail.monologue	0.77	1.31	0.09	−16.8	184	0.00
	5.Mediated.detail.monologue	1.92	0.89	0.06			
Pair 7	7.Actual.inference.monologue	0.95	1.40	0.10	−15.6	184	0.00
	7.Mediated.inference.monologue	2.01	0.87	0.06			
Pair 10	10.Actual.inference.dialogue	0.92	1.38	0.10	−15.3	184	0.00
	10.Mediated.inference.dialogue	1.94	0.90	0.06			
Pair 12	12.Actual.detail.dialogue	1.20	1.47	0.10	−12.6	184	0.00
	12.Mediated.detail.dialogue	2.09	0.91	0.06			

* SEM: Standard error mean

meaning in an independent and self-governing way (Hidri, 2014, 2017). Further, DA attends to the development and learning and supports to discover the test-takers' developing capabilities that are different from their own actual skills (Shrestha & Coffin, 2012). However, as argued by Anton (2009), educators could misrepresent the test-takers' capabilities if they rely on only traditional assessments. Findings of the study were also reported in similar other studies (as an example for this, please check Ahmadi & Barabadi, 2014; Anton, 2009; Haywood & Lidz, 2007; Hidri, 2014, 2017; Hidri & Pileh Roud, 2020; Poehner & Lantolf, 2005).

A high learning potential score indicated that the test-takers' ZPD level was typically close to their own ZAD when the targeted capability was close to internalization (Kozulin & Garb, 2002). Conversely, a low potential score was evidenced by the test-takers' need for more pre-determined hints and some external assistance to internalize the targeted learning potential, and this was reflected in the fact that the test-takers with a low ability level of learning potential score were more successful in taking advantage of the mediation of pre-determined hints than those with high learning potential score. Findings of this study are congruent with Kozulin and Garb' study (2002). However, to our knowledge, regarding the difference between the numbers of hints applied for each question type of listening, no study was found to directly scrutinize the difference between the aforementioned variables. Finally, the statistically significant difference between actual and mediated scores meant a significantly higher performance in the mediated performance rather the actual one. For instance, the SD values in attitude and function test types were higher than the other question types, indicating that the detail scores, inference, as well as the ones of the main idea items were more or less close to one another, while scores in the attitude and function items were more spread. In addition, the mediated scores in both contexts were more homogenous than the actual ones, thus leading to better improvements and more homogeneity of scores. C-DA efficiency in boosting and improving language development and listening is reported in other works (e.g., Ahmadi Safa & Jafari, 2017; Alavi et al., 2012; Lantolf, 2000, 2007; Lantolf & Aljaafreh, 1995; Lantolf & Poehner, 2011; Poehner, 2009; Sadeghi & Khan Ahmadi, 2011).

Computerized learning and assessment will be the norm especially with the outbreak of the COVID-19 pandemic. All educational institutions will adhere to this hybrid mode of leaning, teaching and assessment, by supporting teachers, students, and other stakeholders. It is in this regard that C-DA, and by extension learning, can be used as a complementary mode of learning, along with the face-to-face one. This hybrid mode of learning and assessment can support teachers in creating a variety of challenging authentic learning and assessment situations that could be delivered via computers, thus relegating the presence of teachers to a significant level. However, despite this great leap in educational technology, still the human and affective touch of the teacher is needed to support and mediate learners.

Overall, this study lends support to the positive impact of C-DA on EFL learners' listening ability. Applying manifold formats of DA procedure, encompassing C-DA in a sociocultural context, is regarded as an essential shifting to another paradigm, that of "teaching to the test movement" (Shohamy, 2001) to a "testing to the teaching movement" whose true objective would be to serve and guide test-takers to learn independently. Mediation can potentially activate past knowledge, raise consciousness, and help to boost active learning. And it is in this regard that the current study should be contextualized.

8 Implications, limitations, and recommendations

Several features of C-DA, namely improving the test-takers' listening ability and providing information about their learning potential scores, could empower the language teachers and material and test designers to use such types of assessment in an interactive and productive way. To this end, C-DA offered both EFL learners hints which engaged them with the appropriate tools to diagnose problems and find remedies to the listening problems and language teachers to understand the test-takers' LP in the question types for which there is no need for a tutorial to reflect language test-takers' ZAD, and those for which a tutorial brings about correct responses to integrate language test-takers' ZPD (Poehner & Lantolf, 2013). Additionally, findings of this study bespoke the support role computer software programs can give language teachers and test developers in implementing the main principles of DA, thus permitting educators to simultaneously assess more than one test-taker dynamically. Since C-DA promotes learners' self-assessment and reassessment, such a procedure motivates and inspires language learners to essentially adhere to the language learning and assessment process. For the learners to be assessed and detect their own progress, curriculum and material developers are recommended to use C-DA so that learners and test-takers will not depend on the teacher, as the only information provider and assessor. This alternative form of assessment can practically help learners to continuously assess and reassess their potential to the extent they reach the level where they can perform autonomously. Having the test-takers to validate the pre-determined hints could have led to a more comprehensive validation of the C-DA instrument. However, we thought that the judgement of the TOEFL teachers in their capacity as experts was enough.

One aspect of the delimitations of this study rested on the use of the interventionist approach, and the main limitation of C-DA was that the chances of co-constructing the ZPD decreased (Poehner, 2008). Lidz and Gindis (2003), like the case with this study, stated that integration arises as the intervention is intertwined with the assessment process to explicate the test-takers' capabilities and assist them to reach a more challenging levels. Kozulin (2003) argues that test-takers' cognitive development mainly depends on mastering these instruments; however, these tools might not function successfully in the absence of a mediator hence the limited aspects of C-DA in making the test-takers reach their full potential. Also, adopting a qualitative approach to see how the participants reacted to C-DA would have enlightened the researchers to draw valid inferences on the efficiency of using computers in assessment.

Since the findings indicated the significant and positive impact of C-DA on improving test-takers' listening ability, EFL learners in similar contexts can positively adjust these helpful aspects of C-DA to enhance their listening ability, and EFL teachers should encourage their learners to be involved in DA activities, either individually, in pairs or in groups. C-DA had a significantly positive impact on listening monologues and dialogues of TOEFL, and perhaps EFL teachers are recommended to implement this mode of assessment in various language courses. Educational decision-makers, as well as teachers, should initiate ways to use and apply C-DA along with traditional standardized tests. However, to garner more generalizable data regarding the effectiveness of C-DA, other language skills and even subskills such as vocabulary needed to be explored using the same type of assessment. Curriculum and material developers are recommended to focus on DA and C-DA and suggest various materials and user-friendly software to be used in class. This study can be replicated on other high-stakes tests, by considering a proportionate number of males and females. This study was carried out among learners within the age span of 20–36 years old; the same study could be done among different age groups to check the probable effect of age on performance. Investigating how the learning potential score results mapped out the learners' ZAD and ZPD should be encouraged as future research venue. It is more likely that researchers in the future will try to prove whether C-DA can lead to individualized or group learning. Proponents of DA claim that this assessment is anchored in the Sociocultural theory of Mind; however, observing how the learners react to what the computer gives as output can make us think that in fact C-DA is an individual assessment initiative probably because the affective side of the mediator is not present.

9 Conclusion

This study could be perceived as a timely contribution to the other works undertaken on the integration of C-DA into learning. As stated by many researchers, C-DA is basically conceived as innovative DA procedures since it empowers educators with tools and strategies to assess more than one test-taker in a dynamic and simultaneous way, thus making of assessment an authentic and socio-cognitive activity that needs to be shared by the different participants. Unlike other studies (e.g., Hidri, 2019) applying C-DA puts forward the opportunity for the test-takers to be assessed and reassessed several times and generates an instant scoring profile of each test-taker. It is worthy of mention that applying C-DA does not suggest the elimination of other types of assessment, such as traditional assessment from the educational system, and it is believed that both C-DA and traditional assessment are naturally complementary rather than contradictory.

Appendix 1

“Dear Test-Takers,

Please read the instruction carefully.

This software is designed to test and help you improve your listening comprehension ability in TOEFL IBT.”

“You will hear two academic lectures and two conversations taking place at the university and you will get 18 questions to answer. By clicking on the PLAY icon, you will be able to listen to the conversation or lecture only once. And then click on the ANSWER THE QUESTIONS button to start answering the questions about each part. If you answer a question correctly in the first place you will get an explanation why for example choice (A) is the correct answer if choose an incorrect answer a HINT will be given to you which is listening to a part of the conversation or lecture once more by clicking on the PLAY icon and you have the chance to try again. If you choose the wrong answer again another HINT will be given to you which is written, and you can try another option. If you will not be able to choose the correct answer, finally the correct answer will be given to you along with an explanation. Then by clicking on the NEXT QUESTION button you can move to the next question. If you answer a question in the first place you will get the whole score, but by using each HINT you will lose a score from total score of that question.”

“You have 4 min to answer each question if you do not answer the question in 4 min you will automatically be moved to the next question.”

“Remember you can receive the HINTS if you click on the TRY button only. If you click on the next button you will not be able to go back to the previous question.”

“Your personal information will be safe with us.”

Thank you for your cooperation.”

Appendix 2

Table 10 Rotated component matrixa actual and mediated scores of listening comprehension items

	Component															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
16.Mediated.purpose..ML	.954															
16.Actual.ppurpose ML	.946															
9. Mediated. main idea. Dialogue		.938														
9. Actual. main idea. dialogue		.938														
14. Mediated..detail.. monologue			.935													
14.ac.detail. monologue			.935													
15. Mediated.detail. monologue				.936												
15. Actual..detail monologue				.922												
7. Mediated.inference monologue					.928											
7.ac.inference. monologue					.922											
6. Actual. detail. monologue						.936										
6. Mediated. detail. monologue						.929										
8.me.inference. monologue							.939									
8. Actual. Inference. monologue							.928									
12.me.detail. dialogue								.934								
12.ac.detail. dialogue								.929								
4. Mediated. main idea. monologue									.928							
4. Actual. main idea. monologue									.904							
13. Actual. Detail. monologue										.940						
13. Mediated. Detail. monologue										.930						

Table 10 (continued)

	Component															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
10. Mediated. Inference. dialogue											.935					
10. Actual. Inference. dialogue											.934					
3. Actual. attitude. dialogue												.928				
3. Mediated.attitude. dialogue												.919				
1. Actual. main idea. dialogue													.931			
1. Mediated. main idea. dialogue													.926			
11.ac.inference. dialogue														.947		
11.me.inference. dialogue														.906		
2. Actual. function. dialogue															.911	
2. Mediated. function. dialogue															.909	
5. Actual. detail. monologue																.913
5. Mediated. Detail. monologue																.906
Extraction method: principal component analysis																
Rotation method: varimax with Kaiser normalization																

Extraction Method: Principal Component Analysis

Rotation Method: Varimax with Kaiser Normalization

Rotation converged in 6 iterations

Abbreviations C-DA: Computerized dynamic assessment; DA: Dynamic assessment; LP: Learning potential; LPS: Learning potential score; SCT: Sociocultural theory; ZPD: Zone of Proximal Development; ZAD: Zone of Actual Development

References

Ableeva, R. (2008). The effects of dynamic assessment on L2 listening comprehension. In J. P. Lantolf & M. E. Poehner (Eds.), *Sociocultural theory and the teaching of second languages.* (pp. 57–86). Equinox. https://doi.org/10.1558/equinox.29293.

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. Retrieved on May 26, 2020: http://ilt.atu.ac.ir/article_1759.html.

Ahmadi Safa, M., & Jafari, F. (2017). The washback effect of dynamic assessment on grammar learning of Iranian EFL learners. *Journal of Language Teaching and Learning*, *7*(1), 55–68. Retrieved on May 26, 2020: https://dergipark.org.tr/en/download/issue-file/16558.

Ajideh, P., Farrokhi, f., & Nourdad, N. (2012). Dynamic assessment of EFL reading: Revealing different aspects of different proficiency levels. *World Journal of Education, 4*(2), 102–111.

Alavi, S. M., Kaivanpanah, S., & Shabani, K. (2012). Group dynamic assessment: An inventory of mediational strategies for teaching listening. *Journal of Teaching Language Skills, 30*(4), 27–58. https://doi.org/10.22099/JTLS.2011.370.

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

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

Buck, G. (2001). *Assessing listening* CUP.

Crook, C. (1991). Computers in the zone of proximal development: Implications for evaluation. *Computers and Education, 17*(1), 81–91.

Dixon-Krauss, L. (1996). *Vygotsky in the classroom: Mediated literacy instruction andassessment* Longman.

Elliott, J. (2003). Dynamic assessment in educational settings: Realizing potential. *Educational Review, 55*(1), 15–32.

Emadi, M., & Arabmofrad, A. (2015). Individual dynamic assessment: An analysis of Iranian EFL learners' listening lomprehension errors. *Theory and Practice in Language Studies, 5*(12), 25–39.

Feuerstein, R., & Feuerstein, S. (1991). Mediated learning experience: A theoretical review. In R. Feuerstein, P. S. Klein, & A. Tannenbaum (Eds.), *Mediated learning experience (MLE): Theoretical, psychosocial and learning implications.* (pp. 3–52). Freund.

Ghahremani, D. (2013). The effects of implementing summative assessment, formative assessment and dynamic assessment on Iranian EFL learners' listening ability and listening strategy use. *Journal of language and Translation, 3*(2), 5, 59–68. Retrieved on May 26: https://www.sid.ir/en/journal/ViewPaper.aspx?id=326251.

Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*CUP.

HashemiShahraki, S., Ketabi, S., & Barati, H. (2015). Group dynamic assessment of EFL listening comprehension: Conversational implicatures in focus. *International Journal of Research Studies in Language Learning, 4*(3), 73–89. https://doi.org/10.5861/ijrsll.2014.955.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*CUP.

Heidari, D. M., & Afghari, A. (2015). The effect of dynamic assessment in synchronous computer-mediated communication on Iranian EFL learners' listening comprehension ability at upper-intermediate level. *English Language Teaching, 8*(4), 14–30. https://doi.org/10.5539/elt.v8n4p14.

Hidri, S. (2014). Developing and evaluating a dynamic assessment of listening comprehension in an EFL context. *Language Testing in Asia, 4*(4), 1–19. https://doi.org/10.1186/2229-0443-4-4.

(refereed, Scopus-indexed) (more than 9000 reads: https://languagetestingasia.springeropen.com/articles/10.1186/2229-0443-4-4/metrics).

Hidri, S. (2017). Specs validation of a dynamic reading comprehension test for EAP learners in an EFL context. In S. Hidri & C. Coombe (Eds.), *Evaluation in foreign language education in the Middle East and North Africa* (pp. 315–337). Springer International Publishing. https://doi.org/10.1007/978-3-319-43234-2_19.

Hidri, S. (2019). Static vs. dynamic assessment of students' writing exams: a comparison of two assessment modes. International. *Multilingual Research Journal, 13*(4), 239–256. https://doi.org/10.1080/19313152.2019.1606875.

Hidri, S., & Pileh Roud, L. F. (2020). Developing and using hints in computerized dynamic assessment of a TOEFL iBTreading exam. *Heliyon*, *6*(9), e04985. https://doi.org/10.1016/j.heliyon.2020.e04985.

Jacobs, E. L. (1998). KIDTALK: A computerized language screening test. *Journal of Computing in Childhood Education, 9*(2), 113–131.

Kozulin, A. (2003). Psychological tools and mediated learning. In *Vygotsky's educational theory in cultural context* (pp. 15–38).

Kozulin, A., & Garb, E. (2002). Dynamic assessment of EFL text comprehension of at-risk students. *School Psychology International, 23*, 112–127.

Lantolf, J. P. (2000). Second language learning as a mediated process. *Language Teaching, 33*(02), 79–96.

Lantolf, J. P. (2007). Sociocultural source of thinking and its relevance for second language acquisition. *Bilingualism: Language and Cognition, 10*(1), 31–33.

Lantolf, J. P., & Aljaafreh, A. (1995). A second language learning in the zone of proximal development: A revolutionary experience. *International Journal of Educational Research, 23*(7), 619–632.

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development* OUP.

Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research, 15*(1), 11–33.

Lidz, C. S. (1995). Dynamic assessment and the legacy of L S. Vygotsky. *School Psychology International, 16*(2), 143–153.

Lidz, C. S., & Gindis, B. (2003). Dynamic assessment of the evolving cognitive functions in children. In *Vygotsky's educational theory in cultural context* (pp. 99–116).

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language testing, 18*(4), 333–349.

Mendelsohn, D. J. (1994). *Learning to listen: A strategy-based approach for the second language learner* Dominie Press.

Nunan, D. (1998). Approaches to teaching listening in the language classroom. *Paper presented at the Korea TESOL Conference, Seoul*.

O'Malley, J. M., Chamot, A. U., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition. *Applied Linguistics, 10*(4), 418–437.

Peña, E., Iglesias, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10*(2), 138–154.

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

Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *Modern Language Journal, 91*(3), 323–340.

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting second language development* Springer.

Poehner, M. E. (2009). Group dynamic assessment: Mediation for the L2 classroom. *TESOL Quarterly, 43*(3), 471–491.

Poehner, M. E., & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research, 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: An International Journal, 17*, 312–330.

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

Poehner, M. E., Zhang, J., & Lu, X. (2015). Computerized dynamic assessment: Diagnosing L2 development according to learner responsiveness to mediation. *Language testing, 32*(3), 337–357.

Rost, M. (2002). *Teaching and researching listening* Longman.

Sadeghi, K., & Khan Ahmadi, F. (2011). Dynamic assessment of L2 grammar of Iranian EFL learners: The role of mediated learning experience. *International Journal of Academic Research, 3*(2), 931–936.

Scarcella, R. C., & Oxford, R. L. (1992). *The tapestry of language learning: the individual in the communicative classroom* Heinle & Heinle.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests* Longman.

Shrestha, P., & Coffin, C. (2012). Dynamic assessment, tutor mediation and academic writing development. *Assessing Writing, 17*(1), 55–70.

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. *Journal of Computer Assisted Learning, 18*(1), 21–32.

Underwood, M. (1989). *Teaching listening* Longman.

Vandergrift, L. (2004). Listening to learn or learning to listen. *Annual Review of Applied Linguistics, 24*, 3–25.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching, 40*(3), 191–210.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*Harvard University Press.

Vygotsky, L. S. (1986). *Thought and language* MIT Press.

Wang, P. (2015). The Effect of dynamic assessment on the listening skills of lower- intermediate EFL learners in Chinese technical college: A pilot study. *Journal of Language Teaching and Research, 6*(6), 1269–1279.

Wang, T. H. (2008). Web-based quiz-game-like formative assessment: Development and evaluation. *Computers & Education, 51*(3), 1247–1263.

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