The potential of an adaptive computerized dynamic assessment tutor in diagnosing and assessing learners' listening comprehension

Izadi, Mehri; Izadi, Maliheh; Heidari, Farrokhlagha

DOI
10.1007/s10639-023-11871-w

Publication date
2023

Document Version
Final published version

Published in
Education and Information Technologies

Citation (APA)
Izadi, M., Izadi, M., & Heidari, F. (2023). The potential of an adaptive computerized dynamic assessment tutor in diagnosing and assessing learners' listening comprehension. *Education and Information Technologies*, *29 (2024)*(3), 3637-3661. https://doi.org/10.1007/s10639-023-11871-w

Education and Information Technologies (2024) 29:2303–2327
https://doi.org/10.1007/s10639-023-11871-w

The potential of an adaptive computerized dynamic assessment tutor in diagnosing and assessing learners' listening comprehension

Mehri Izadi[1] · Maliheh Izadi[2] · Farrokhlagha Heidari[1]

Received: 12 August 2022 / Accepted: 2 May 2023 / Published online: 24 June 2023

Abstract

In today's environment of growing class sizes due to the prevalence of online and e-learning systems, providing one-to-one instruction and feedback has become a challenging task for teachers. Anyhow, the dialectical integration of instruction and assessment into a seamless and dynamic activity can provide a continuous flow of assessment information for teachers to boost and individualize learning. In this regard, adaptive learning technology is one way to facilitate teacher-supported learning and personalize curriculum and learning experiences. This study aimed to investigate the potential of an adaptive Computerized Dynamic Assessment (C-DA) tool applicable as a language diagnostician and assistant. The study tried to get insight into 75 Iranian EFL learners' listening development by focusing on the learning potential exhibited through learners' assessment and the degree of internalization of mediation. To achieve these, a C-DA tutor including two dynamic listening comprehension tests, each comprising 20 items, arranged in the order of difficulty was developed. The test takers unable to answer an item correctly were provided with graduated hints for different comprehension- and production-type items and the overall difficulty level of the test was adapted to the test takers' proficiency level. In order to have a full diagnosis of each individual's listening development, the adaptive C-DA automatically generated five test scores on each learner's performance: actual (unmediated) score, mediated score, gain score, Learning Potential Score (LPS), and transfer score. The results of paired-sample t-tests revealed a significant development from the actual to the mediated scores. Furthermore, the LPSs indicated that the tutor was capable of revealing learners' potential for learning. Moreover, learners with high LPS gained a higher mean for transfer scores followed by transfer scores of medium and low levels. The results of Mann-Whitney tests revealed a significant difference in the degree of internalization of mediation of learners with mid and low range of LPSs on the easy test and high and low range of LPSs on the difficult test. The findings of this research can have important theoretical and

Extended author information available on the last page of the article

practical implications for researchers and educationalists. The instructional value of this adaptive C-DA tool lies in its unique opportunities for individualizing learning and developing individual learning plans in accordance with learners' needs.

Keywords Dynamic assessment · Computerized dynamic assessment · Adaptive computerized dynamic assessment tutor · Learning potential · Internalization of mediation

1 Introduction

Recently, Learning-Oriented Assessment (LOA) has emerged as an alternative assessment procedure due to the gush of interest in diagnostic language assessment and the necessity of identifying the learners' sources of problems. The term learning-oriented assessment has been widely used to "encapsulate the idea that all forms of assessment should promote learning" (Green, 2016, p. 23). In LOA, the focus shifts from goal-driven to learning-driven assessment which can be viewed as assessment as learning (May et al., 2019). The recent diagnostic language assessment needs more than diagnosis-making instruments and entails a diagnostician who can make a diagnosis, a user-friendly, purpose-built system that allows for diagnostic stages, and a structure that includes diverse stakeholders' views and provides feedback for subsequent treatment (Alderson et al., 2015; Harding et al., 2015).

Alderson et al.'s (2015) theory of diagnostic assessment theorizes the multifaceted nature of diagnosis in language assessment and typifies diagnostic testing procedure as a distinguishing feature to help plan individualized schemes of work for each learner. Aligning with Alderson et al.'s (2015) theory of diagnosis, Dynamic Assessment (DA) with its reliance on teacher-learner dialogue during the assessment, allows diagnosing of specific sources of difficulty, giving insights into the process of learning, providing purposeful information, and observing language development. The Vygoteskian notions of the Zone of Proximal Development (ZPD) and Zone of Actual Development (ZAD) are central to DA. ZPD is defined as the gap between what a learner can perform independently and what he/she can do under cooperation with a more knowledgeable person (Haerazo et al., 2019). ZAD, on the other hand, reflects the learners' independent and unassisted performance which can function as a diagnosis of what is currently needed to move to ZPD (Bakhoda & Shabani, 2018). Mediation is believed to play a central role in diagnosing the learners' underlying capabilities in terms of both unassisted (ZAD) and assisted (ZPD) performances. Although it may seem uncertain how to be accomplished best, both educators and psychometricians put emphasis on the necessity of bringing instruction and assessment into a closer nexus. Consolidating assessment and instruction into a unified activity seemed not difficult to Vygotsky due to the dialectic nature of his theory. To Vygotsky, instruction and cognitive development create a dialectical unity in the sense that instruction paves the way for development to go forth and development brings up opportunities for future instruction (Poehner & Lantolf, 2005). In this way, a true diagnosis discloses the areas of difficulties and presents a prognosis for overcoming them. In other words, a complete assessment demands instruction which roots in

Vygotsky's argument that higher cognitive development finds its way through sociocultural activity in the form of mediation or appropriate assistance. In DA, a mediator/assessor provides intentional and reciprocal support to the learner when s/he faces difficulties during the assessment procedure. Generally, DA studies have adopted one of the two general approaches to promoting ZPD development (Ahmadi Safa & Beheshti, 2018). Some of them have taken a psychometric orientation using a set of standardized menus of hints and clues. This approach follows Vygotsky's early writing on IQ testing. Lantolf and Poehner (2004) call it as interventionist approach to DA which is informed by Vygotsky's quantitative interpretation of the ZPD as a 'difference score'. This approach is more formal and standardized and better lends itself to quantification and statistical analysis. However, since the mediator standardizes the mediation, it is similar for all learners (Fulcher, 2010). On the other hand, others may take a dialectic approach which is more in tune with Vygotsky's (1998) understanding of the diagnosis. This orientation which is called the interactionist approach (Lantolf & Poehner, 2004) is informed by Vygotsky's qualitative interpretation of the ZPD and encompasses an open-ended qualitative collaboration between mediator and learner (Poehner, 2008). Leading questions and prompts are not planned ahead but they arise from the mediated dialogue. In fact, this collaboration justifies the use of the word dynamic (Fulcher, 2010).

However, there are some frequently reported problems with DA including its time-consuming nature in one-to-one interactions which can be alleviated through the deployment of other forms of DA such as the Computerized Dynamic Assessment (C-DA) format (Poehner et al., 2015). Likewise, Alderson et al. (2015) point out that a theory of diagnosis should not preclude large-scale assessment and emphasize that appropriate assessment tools can even pose a challenge to such programs. Poehner et al. (2015) mention the concurrent administration to a large group of individuals, the desired self-assessment and re-assessment of learners, and the spontaneous production of reports on each person's performance as advantages not possibly attainable through non-computerized DA. Providing mediation through e-mails (Shrestha & Coffin, 2012), Web-based DA system (Wang, 2010), Web 2.0 (Birjandi & Ebadi, 2012), and reading computerized DA (Bakhoda & Shabani, 2018; Hidri & Pileh Roud, 2020; Shabani, 2012; Yan & Qian, 2020) revealed that the C-DA format creates "an assessment-centered e-learning environment that treats assessment as teaching and learning strategy" (Wang, 2010, p. 1165). This can be achieved by juxtaposing mediation and feedback to permit a larger group of individuals to investigate or utilize some essential principles so that they could independently solve problems and accordingly learn more. The recent attempt to conduct C-DA follows the interventionist approach to DA, in which the mediation is not tuned to each individual's learning needs (e.g., Birjandi & Ebadi 2012; Wang, 2010). In other words, all individual learners receive a pre-specified set of hints for all items irrespective of item construct (e.g., phonetics, pragmatics) and item mode (e.g., comprehension & production). Accordingly, learners' strong and weak points in the tested areas cannot be traced (Shabani, 2012).

The existing problems with the current C-DA motivated the researchers to develop an adaptive C-DA tutor with the capability of providing diagnostic feedback and remedial instruction to learners, a learning-oriented assessment tool rather than a test-

ing tool. The mediation embedded in adaptive C-DA would be prepared according to an interactionist approach to DA in which open-ended mediator-learner dialoguing during joint activity leads to mediation with no constraints (Poehner, 2005).

The possibility of learners' listening comprehension development through DA in English as a Foreign Language (EFL) context has been investigated by Ableeva (2010), Khoshsima and Izadi (2014), Barabadi et al. (2018), and Heidari and Izadi (2020). While Ableeva (2010) examined the effect of listening DA on 7 French learners through the interactionist approach, Khoshsima and Izadi (2014) investigated the interventionist approach to DA in the forms of dynamic-supported and dynamic-instructed assessments on 38 EFL learners. Barabadi et al. (2018) took a step forward and integrated the interactionist and interventionist DA approaches in assessing learners' listening comprehension. Similarly, Heidari and Izadi (2020) explored DA potential in developing the listening comprehension ability of 453 Second Language (L2) learners in a listening comprehension Massive Open Online Course (MOOC) using a microgenetic approach. In other words, to prepare appropriate hints for each listening item, they utilized an interactionist DA while an interventionist DA was employed for providing feedback. Although, in these studies, the efficacy of DA was assessed in the form of mediator-learner interaction, the C-DA was not considered. Poehner and Lantolf (2013) and Poehenr et al. (2015) assessed learners' comprehension in C-DA format. However, they could not catch a complete picture of the test takers' ZPD. More recently, Kao and Kuo (2021) investigated the effectiveness of C-DA to identify the listening difficulties of L2 learners, diagnose their specific learning needs, and assess their prospective potential abilities through their performance on 3 question types: overview, detail, and inference questions in multiple-choice format. The C-DA program proved effective in offering a fuller picture of learners' listening abilities by automatically generating scores on each individual's current abilities, mediated performance, learning needs, and future potential. It is worth mentioning that Kao and Kuo (2021) included just comprehension-type items and the provided mediation was not sensitive to test takers' listening proficiency level. Therefore, the potential of an adaptive C-DA capable of providing mediation on comprehension- and production-type items and sensitive to learners' language proficiency level for a large group of EFL learners' listening comprehension development has not yet been investigated. Accordingly, the goal of this study is to investigate whether ZPD-sensitive instruction offered in DA through juxtaposing assessment and instruction is similarly applicable to an adaptive C-DA format attuned to the individual needs of a larger number of learners. ZPD-regulatory hints are believed to assess learners' actual development, uncover their emerging abilities, and reveal a larger picture of learners' development by depicting a potential future (Poehner & Lantolf, 2013). Since Vygotsky's approach to ZPD is closely related to three interconnected and prominent constructs of development, potential for learning, and internalization of mediation, this study tries to provide answers to the following research questions:

1) Does the adaptive C-DA tutor have the ability to develop EFL learners' listening abilities?
2) Does the adaptive C-DA tutor have the ability to reveal EFL learners' potential for learning?

3) Does the adaptive C-DA tutor have the potential of enhancing the internalization of mediation?

2 Literature review

2.1 Theoretical framework

Dynamic assessment has its roots in Vygotsky's Sociocultural Theory (SCT) which asserts that human cognitive development and learning is a social phenomenon mediated by social interactions. DA is defined by Haywood and Lidz (2007) as "an interactive approach to conducting assessments within the domains of psychology, speech/language, or education that focuses on the ability of the learner to respond to intervention" (p. 1). In fact, DA was introduced to assist and simultaneously assess learners in accordance with what is above their current level of ability or ZAD. Mediation is supposed to play a key role in assisting learners to traverse the distance between ZAD and ZPD, their near-future potentiality. C-DA embraces the typical implementation of interventionist DA through computer software. The sequence and types of mediation are predetermined in a standardized manner through computer algorithms, and psychometric principles are used to quantify change(s) in learners' behavior (Shabani, 2014). A C-DA with its pre-specified mediatory moves can be administered to a large number of individuals and can produce comprehensive profiles which include the test-takers' current developmental levels, their potential developmental levels, and some other information that reveals their microgenetic developments and their potential for genitive modifiability. As Poehner (2008) states, appropriate forms of interaction and instruction can intervene in and bring about a process of progression in the cognitive abilities of human beings which can be viewed as open systems as opposed to their pure genetically determined potentials. Like all other interventionist approaches towards DA, mediations given in C-DA are limited since they may not be calibrated to the ZPD of all learners (Poehner & Lantolf, 2013). Despite these limitations, it assuages many shortcomings of DA by placing mediation and feedback in proximity to permit a larger group of individuals to benefit from mediated support so that they could independently solve problems and correspondingly learn more. As a reaction to traditional or non-adaptive testing, adaptive testing, also referred to as Computer Adaptive Testing (CAT), has attracted educators and researchers' attention in recent decades. In fact, computer technology has made it plausible to adapt the difficulty level of test items to the test taker's proficiency level. In other words, according to an estimation of the current ability level of the test taker, the item believed to provide most information is presented next. This type of assessment comes in contrast to linear assessment in which all candidates answer a set of similar items in a predetermined order irrespective of their performance on prior items (Wang, 2010). Modern CAT is founded on tenets taken from Item Response Theory (IRT) which expresses the probability of a correct answer to an item by juxtaposing the difficulty level of the test items and the testee's answers (Chen & Wang, 2010). In this way, computer technology has paved the way to dynamically adapt the difficulty level of test items to the test taker's real-time performance (Van der Linden & Glas, 2010).

In CAT, the first item is randomly chosen from an item pool and the test taker's performance on this item dictates the succeeding items (Veldkamp & Matteucci, 2013). The available literature mentions some distinguishing features as the advantages of CAT over traditional testing. Wainer (2000) claims that the deployment of CAT can significantly reduce test administration time by excluding those items which are too easy or too difficult for learners. Therefore, items are presented in a way to reveal the candidates' maximum level of performance (Dörfler et al., 2017). Automated scoring and large item pools which facilitate the development of multiple versions of a test with different difficulty levels are still some other advantages of CAT.

The literature on CAT also distinguishes between selected-response and constructed-response items. The former includes items with existing answers to be chosen by examinees, while the latter comprises open-answering items (Ward & Bennet, 1993). The currently available and deployed CAT systems mainly comprise items that can be answered dichotomously or on a multi-part scale (Oppl et al., 2017). However, the evaluation of learning outcomes usually demands more complex and open-answer items (Guzmán & Conejo, 2005).

2.2 Empirical studies

In the realm of the application of C-DA, most studies have deployed the interventionist approach using pre-specified mediations. To mention some, Pishghadam et al. (2011) designed a C-DA for testing reading comprehension and evaluated its prefabricated hints and mediation effectiveness. The comparison of learners' ZAD and ZPD scores supported the effectiveness of the software in improving learners' reading comprehension. Ebadi and Saeedian (2019) explored the C-DA effectiveness in realizing learners' latent potential in reading comprehension. The results proved the software's effectiveness in diagnosing learners' potentials since learners with almost similar pre-test scores turned out to obtain different or even highly different DA scores on post-test and, therefore, different Learning Potential Scores (LPSs). More recently, focusing on learners' pragmatic knowledge, Zangoei et al. (2019) deployed a web-based C-DA and an interventionist approach towards DA in an attempt to integrate assessment and instruction. The results indicated the improvement of test takers' pragmatic knowledge as well as different levels of responsiveness to mediation between individuals in line with their different ZPD levels. Similarly, Mehri Kamrood et al. (2021) designed and utilized an online multiple-choice C-DA software for testing EFL university students' listening skills. Based on an interventionist approach, three sets of scores including actual, mediated, and LPSs were provided. They reported a significant difference between learners' actual and mediated scores; therefore, the software proved efficient in diagnosing the learners' weaknesses and strengths in different language constructs.

However, Poehner and Lantolf (2013) and Poehner et al. (2015) endeavored to remedy the downsides of earlier C-DA by integrating the interactionist and interventionist models of DA to prepare mediations. They collected data through interactionist DA and prepared a standardized menu of hints/prompts for each individual item based on interventionist DA. In these two studies, they designed online multiple-choice tests of L2 reading and listening comprehension in Chinese, Russian, and French.

Following the interactionist approach to DA, these studies qualitatively collected and scripted a list of prompts. Then, in an interventionist approach to DA, a standardized menu of prompts was produced to be shown for each individual item. When learners failed to respond to an item correctly, 4 prompts were provided arranged from the most implicit to the most explicit. Similarly, Yang and Qian (2020) explored the potential of C-DA in assessing Chinese college students' reading comprehension through the integration of interventionist and interactionist approaches. Both Poehner and his co-authors and Yang and Qian (2020) concluded that the C-DA was able to diagnose individuals' independent and mediated performances and to track their improvement and development through learning potential and transfer scores. Although these researchers attempted to represent hints/prompts which equated with the type of mediation, the mediation was not sensitive to test takers' ZPD. The adaptation of test takers' ability to the difficulty level of the test and the adjustment of mediation to the item construct and item mode were also missing.

Most studies on CAT have compared the efficiency of computerized adaptive versions of tests being available. For instance, Burston et al. (2016) designed a computerized adaptive test based on the existing Saint Louis University Placement Exam (SLUPE) and compared the placement results with testees' self-evaluations and instructor ratings. The close correlation between the computerized adaptive test and instructor ratings convinced them to base their diagnostic decisions on the computerized version of the test. Similarly, Tseng (2016) examined the potential of CAT in estimating the English vocabulary size of test takers. He also compared the efficiency and accuracy of the computerized version with the original paper and pencil test and whether they were equally able to categorize testees into mastery and non-mastery groups. It was concluded that only one-third of the items in the item bank were required to estimate the participants' vocabulary size. The results also confirmed the precision and efficiency of the CAT in estimating the vocabulary size of the examinees and classifying them into mastery and non-mastery groups. More recently, Mizumoto et al. (2019) designed a CAT version of the Word Part Levels Test (WPLT) with the aim of boosting the diagnostic features of the original test. The comparison of the results of the WPLT and its CAT version indicated that the computerized adaptive version of the test could supply more diagnostic information in terms of each candidate's strengths and weaknesses in affix knowledge based on a smaller number of items but with the same or even greater precision. However, the obtained results of these studies were based on learners' ZAD or unassisted performance and candidates received no mediation.

As the available literature indicates, the mediation in the aforementioned DA studies is not sensitive to test takers' ability level, the type of construct assessed, and the mode of presenting an item. Besides, the deployed adaptive tests do not reveal the type and degree of support and mediation required by an examinee to answer each particular item. The confluence between DA and CAT may provide more information on a larger number of learners' performance including the efficiency of mediation in ZPD development, revealing the potential for learning, and the internalization of mediation. Therefore, the present study aimed to fill the existing gap by developing an adaptive C-DA functioning as an auxiliary tool with the aim of developing learners' listening comprehension. In so doing, an adaptive C-DA was designed and

its effect on EFL learners' listening comprehension development and its ability in revealing the learners' learning potentials and enhancing the degree of internalization of mediation were explored.

3 Methodology

3.1 Framework and design of the study

The theoretical underpinnings that form the framework of this research include the diagnostic assessment theory outlined by Alderson et al. (2015), Vygotskian SCT underlying DA, and IRT as the foundation of adaptive assessment. Alderson et al. (2015) were among the first who tried to propose a tentative framework for diagnostic assessment applicable to Second and Foreign Language (SFL). Theorizing diagnosis as a forerunner of further intervention demands information on the actual cause of the problem rather than a symptom description. DA has its roots in SCT, expanded by Vygotsky, stating that fine-grained mediation and each learner's responsiveness to mediation can trigger future development. With regard to adaptive assessment, this study is grounded in IRT which strives to juxtapose item characteristics and learner characteristics to assert correctly answering the item. As Chen and Wang (2010) state, IRT tries to calibrate each item to each individual's skill level. A variety of parameters such as item difficulty, discrimination (the amount of information each item provides for estimating each individual's skill level), and the selected-response or constructed-response item can form the basis of this calibration.

A single-group experimental design with the intervention was deployed to map out a quantitative capture of learners' performance, learning potentials, and internalization of mediation by reporting a scoring file for each individual.

3.2 Participants

The target population of this study included groups of Iranian EFL learners (n = 170) enrolling in an online International English Language Testing System (IELTS) course. Information on the demographic characteristics of the participants including name, age, gender, educational degree, and place of living was collected at the enrollment time. The learners' age ranged between 18 and 32 and they came from different cities and universities in Iran. Among these, a sample comprising 75 learners (35 males & 40 females) was randomly selected to take the C-DA test. To ensure that the sample was representative of the target population and to minimize bias, stratified random sampling was deployed. In other words, the participants were divided into subgroups, or strata, in terms of similar attributes including gender, age, and educational degree, and then random sampling was done from each subgroup. The link to the C-DA program package was shared through e-mail with the participants. The prospective participants were provided with pertinent information to make an informed consent to participate in the study and they were ensured that their data would be kept confidential. These 75 learners took the C-DA test and sent back their scoring files to one of the researchers' email address.

3.3 Instruments

The instrument is a C-DA program package with the capability to dynamically test the listening ability of the participants by presenting predetermined prompts if their answers were incorrect. This tool is a desktop .NET application written in C# programming language with Visual Studio using the WPF (Windows Presentation Foundation) user interface framework. The adaptive C-DA uses the MVVM (Model-View-ViewModel) architectural pattern to separate the view from the model. The view defines the structure, layout, and appearance of the app for users to see and interact with. The view model implements data binding commands and notifies the view about state changes. The model classes contain the app's data, business, and validation logic. When a participant clicks a button in the view of the app, a command on the view model executes to perform the requested action. The adaptive C-DA's logic is rule-based, i.e., it follows predetermined rules based on the expertise of the test takers.

The adaptive C-DA consisted of three parts: the introduction section, the main part, and the learner's profile scores. The first part was an introduction including a short description of DA and comprising some questions on learners' personal characteristics including name, age, gender, etc. The second part included two dynamic listening comprehension tests arranged in the order of difficulty. Each test comprised 20 items, and each item was followed by hints for the participants unable to answer it correctly. The third part presented a scoring file of the participant's performances. The scoring file was able to present each learner's actual and mediated scores, the number of hints used based on item construct and item mode, and the total time used to answer the test.

3.4 Procedure

3.4.1 Test preparation

The following considerations were taken into account when selecting the listening tests for the study. First, the areas in which EFL learners had comprehension difficulties were pursued. Focusing on the sources (i.e., phonology, lexis, syntax, context, & culture) which were problematic in listening comprehension, 58 listening items were extracted from the Cambridge IELTS books 1 to 9. The listening items were of comprehension (n=31) and production (n=27) formats. The comprehension items included multiple-choice, matching, and choosing a word from a list; the production items included short answers, completing a picture or table, filling in gaps, and summary completion.

To better serve the purpose of a DA test and to give learners the chance of multiple attempts and mediation to answer each item, one additional distractor was added to each item (e.g., bringing the total number of choices per item to 5). Test piloting helped to specify item characteristics after the changes were made. To assure the content validity of the test before piloting, it was reviewed and judged by three testing experts. They qualified the content of the test to be appropriate for measuring the listening ability of the learners.

Table 1 The Descriptive Statistics and Reliability Level of the Easy and Difficult Tests

	Mean	Standard Deviation	Reliability (Cronbach's alpha)
Easy Test	15.86	2.04	0.898
Difficult Test	14.04	1.72	0.850

Table 2 Correlation between the Scores on the Listening Section of the FCE Test and Easy Test

		FCE
Easy	Pearson Correlation	0.775
	Sig. (2-tailed)	0.037
	N	121

*Correlation is significant at the 0.05 level (2-tailed)

3.4.2 Test piloting

In order to specify the reliability and validity level of the test and item characteristics, the 58 listening items were distributed among 121 EFL learners. Since the developed C-DA was to be adaptive in the sense that it would consist of two subtests arranged in the order of difficulty, it was important to know the difficulty level of items. Accordingly, based on the specified difficulty level, the items were divided into two easy and difficult subtests. Items with the difficulty level of 0.62 and above were categorized as the easy test and those with the difficulty level of 0.32 and below as the difficult one. Moreover, in order to make sure that these two tests were adequately different from each other in terms of difficulty, items with difficulty levels between 0.32 and 0.62 were omitted (18 items). Table 1 represents the reliability of the tests computed through Cronbach's alpha formula.

As Table 1 indicates the reliability indices of the two tests were rather high.

In order to validate the tests, it was necessary to demonstrate a relationship between the scores on these tests and the scores on other standardized tests believed to be indicators of the ability being tested (Bachman & Palmer, 2010). The standard tests chosen as the criteria for the validation process were the listening section of First Certificate English (FCE) for the easy test and Advanced Certificate English (ACE) for the difficult test published by Cambridge University. The rationale for selecting these tests was that these tests, like the IELTS test, could measure different real-life listening skills. The Pearson product-moment correlation formula was applied to estimate the correlation between the standard tests and the devised easy and difficult tests. As illustrated in Table 2, the correlation coefficient (i.e., criterion-related validity) between FCE and the easy test was 0.77 indicating that there was a fairly high correlation between these two tests.

Table 3 reveals the correlation coefficient (i.e., criterion-related validity) between ACE and the difficult test. A reliability value of 0.89 indicates a fairly high correlation between these two tests.

Table 3 Correlation between the Scores on the Listening Section of the ACE Test and Difficult Test

		ACE
Difficult	Pearson Correlation	0.899
	Sig. (2-tailed)	0.033
	N	121

*Correlation is significant at the 0.05 level (2-tailed)

3.4.3 Designating hints

The investigation of the DA approach ran in a one-on-one tutoring format. The mediational moves and learner reciprocity for all listening items were extracted and the standardized menus of mediating moves were scripted to accompany each listening item. The process of designating hints benefited from both interactionist and interventionist DA. In other words, the scripted mediating hints in the computerized version of the tests presented an interventionist model of DA but the hints themselves were extracted from the interactionist DA based on pilot administrations of the tests. In this way, after meticulous analysis of mediational moves and learner reciprocity, the menus of mediating moves were scripted for each item.

The mediator-learner interactions were further analyzed. The analysis revealed that the mean number of hints learners received to answer an item correctly was 4 for comprehension items and 6 for production items. This study examined the number of prompts presented to the learners in order to help them complete an item not based on prior decisions of the researchers. With respect to the production-type items, it is worth noting that if the menus of hints had more than 6 hints, it might make the learners bored and less motivated in completing the test. Whereas, if the menus of the hints had less than 6 hints, it might not allow learners to demonstrate their potential.

It is worth mentioning that the hints were designed with respect to the sources of listening comprehension problems and item mode. Moreover, while the precise content of the hints differed across items, each followed the route moving from the most implicit to the most explicit hint. If a learner's response was incorrect, the most implicit mediating prompt was provided and there was a chance to reattempt the item. Maximally 4 and 6 mediating prompts were included for comprehension and production items, respectively; each becoming more explicit until finally the correct answer was chosen and an explanation for the solution was offered.

More importantly, each hint presented two kinds of help. The replaying move (replaying the entire part/replaying a segment from the part) turned out to be the most frequent mediational move used by the mediator and the requesting move to replay was the most frequent response presented by the learners. Moreover, it was revealed that listening to the text for the second or third time could be more helpful in answering an item than necessarily providing a prompt.

3.4.4 Program preparation

As a tutor, the adaptive C-DA was equipped to mediate learners to improve their listening comprehension skills through providing graduated hints for different comprehension- and production-type items.

First, the easy test consisting of 20 items was displayed. If learners were able to answer an item correctly, they would move on to another item. However, if they answered an item incorrectly, predetermined hints arranged from the most implicit to the most explicit were shown. The number of hints used for the first five questions of the easy test including 3 comprehension and 2 production items helped to estimate the proficiency level of the learners. That is whether the learners should continue with the easy test or be directed to the more difficult one. On average, if a given learner used 12 hints or less, the test was considered easy for him/her, and s/he would be directed to the second test which was more difficult. In other words, for learners whose average use of hints was two or below for each comprehension item and three or below for each production item; the test was considered within their ZAD. Therefore, they were directed to the second test which was more in accordance with their ZPD. This partial adjustment of the test takers' ability to the difficulty level of the test could somehow obviate one of the main deficiencies of C-DA, namely the lack of sensitivity of mediation to test takers' ZPD.

Moreover, the adaptive C-DA was capable of adjusting to participants' language proficiency levels. First, based on a participant's performance on the first 5 questions, each individual's listening ability was estimated. Then, during the next 15 questions of the easy test, the program actively tried to adjust the next hints based on the weaknesses of each individual evident in the answers to previous questions. That is, the history of learners' answers (their interactions with the program) was also a factor in our model. For instance, by the second try, learner A might receive Hint #2 which helps him/her focus on pronunciation, while learner B might receive Hint #3 which alerts the learner regarding structure. In doing so, the researchers were able to make the learners more conscious of their weak points and accordingly provide hints. The main goal here was to provide a program that was adaptive enough to provide hints that were tailored to individual needs of learners. Figure 1 displays how the program functions.

Upon preparing the computerized dynamic test, it was piloted with 35 EFL learners to explore the hints' effectiveness. Accordingly, the hints were reanalyzed and some modifications were made to make them more understandable and more harmonized with the test takers' ZPD. Finally, the final version of the test comprising the items and hints was reviewed by two experts and some minor changes were made.

3.5 Data analysis and coding/scoring procedure

In order to have a full diagnosis of an individual's listening development, the adaptive C-DA automatically generated five test scores on each learner's performance. In what follows, these five score types are introduced.

- The actual (unmediated) score indicated an individual's independent, unmediated performance, ZAD. For production test items, if a learner responded correctly on the first attempt, s/he would receive the maximum point (6) and if s/he responded incorrectly, s/he would receive the minimum point (0). For comprehension test items, if a learner responded correctly on the first attempt, s/he would receive

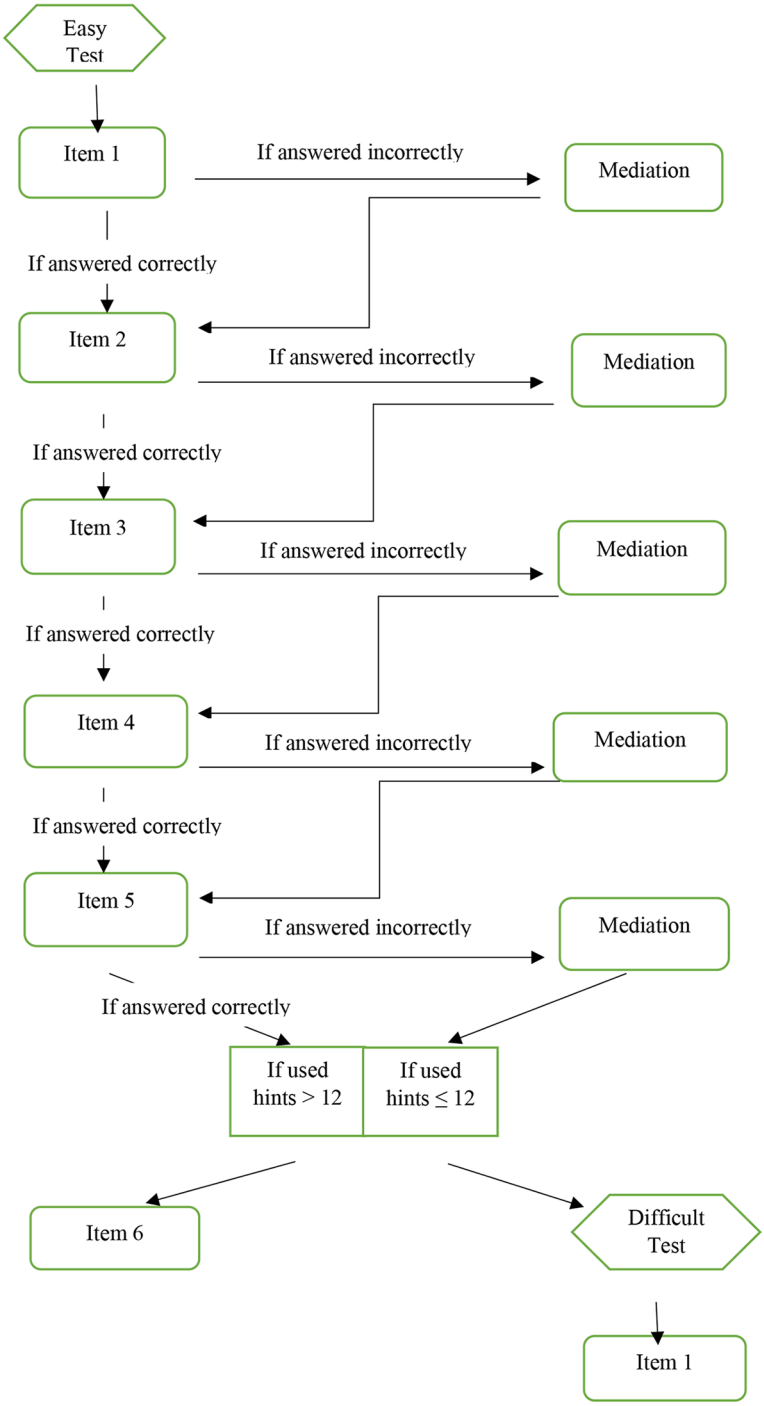

Fig. 1 The Flowchart of Easy and Difficult Tests

the maximum point (4) and if s/he responded incorrectly, s/he would receive the minimum point (0).

- The mediated score revealed an individual's mediated performance, ZPD. For each test item answered incorrectly, a series of hints were shown until the learner answered correctly or the correct response was provided by the program. The mediated score indicated the number of mediating prompts each learner received, the total point granted for that item was decreased by one. For example, if the learner's first response to a production item was correct, a score of 6 would be granted for that item. If the learner's second attempt at the same item produced a correct response, a score of 5 would be granted, and so on until the correct answer was indicated or selected by the learner and a score of 0 would be assigned for that item. The same happened to comprehension items. For example, if the learner's first response to that item was correct, a score of 4 would be granted. If the learner's second attempt at the same item produced a correct response, a score of 3 would be granted, and so on until the correct answer was indicated and a score of 0 was assigned. Then, for any given production item the learner's mediated score might be any number ranging from 0 to 6 and for any given comprehension item the learner's mediated score might be any number ranging from 0 to 4 based on whether and how much mediation was provided.
- The gain Score (GS) showed the amount of change between the unmediated and mediated performances.
- The Learning Potential Score (LPS) revealed an individual's level of improvement and progress under mediation. The LPS is calculated through the following formula.
  LPS= (2 * Mediated Score − Actual Score) / Maximum Score
  According to Kozulin and Garb (2002), LPS score is divided into three levels:

  1. LPS ≥ 1.0 as high.
  2. 1.0 > LPS ≥ 0.71 as medium.
  3. LPS < 0.71 as low.

- Transfer score indicated an individual's ability to transfer the newly acquired knowledge to novel contexts. In this case, it demonstrated the individual's ability to use the newly acquired knowledge to answer more difficult listening items. The listening test included two types of items: transfer and non-transfer items. The transfer and non-transfer items were constructed and tested with the same number of response options, hints and focus on a particular model and sub-skills relevant to comprehension and production. The difficulty level was the only difference between transfer and non-transfer items.

Table 4 provides a summary of different test scores generated by the adaptive C-DA program and what each of them presents.

Table 4 Different Test Scores Generated by the Adaptive C-DA Program

Actual (unmediated) score	indicates the independent, unmediated performance, ZAD
Mediated score	reveals the mediated performance, ZPD
Gain Score (GS)	shows the amount of change between the unmediated and mediated performances
Learning Potential Score (LPS)	reveals an individual's level of improvement under the mediation
Transfer score	indicates an individual's ability to transfer the newly acquired knowledge to novel contexts

Table 5 Descriptive and Inferential Statistics of Non-DA, DA, and Gain Scores for the Easy Test

	N	Mean	SD	t (*p*)
Non-DA score	31	15.41	6.89	14.04 (0.000)
DA score	31	46.77	9.06	
Gain score	31	31.35	12.42	

4 Results

4.1 Research question 1

The first research question aimed to investigate the feasibility of developing the listening ability of EFL learners through the adaptive C-DA. To provide a plausible answer to this question, the learners' actual and mediated scores on the two tests were statistically processed. Out of 75 participants, 31 took the easy test. In other words, these 31 learners used more than 12 hints for answering the first five items. Therefore, they continued with the 15 remaining items of the easy test which was close to their ZPD. The rest of the participants, 44, used 12 or fewer hints, meaning that the first test was within their ZAD. Hence, they were directed to the more difficult test which was within their ZPD. Comparisons of the mean scores of the actual and mediated scores of the 31 test takers who took the easy test showed a change of mean scores from 15.41 (SD=6.89) to 46.77 (SD=9.06). The results of paired-sample t-test are presented in Table 5.

Table 5 reveals that this difference between the actual and mediated scores of test takers on the easy test was significant (t (30)=14.04, p<. 01, d=3.89). It is worth mentioning that a significant and high correlation (r=.70, p=.00) was found between actual and mediated scores of the learners on the easy test, indicating a direct mapping across scores. The results also showed a strong negative correlation between gain scores and actual scores (r=−.69, p=.00) indicating that learners who did not perform well independently benefited more from mediation than those who performed better independently.

Table 6 displays the descriptive and inferential statistics of learners who took the difficult test. Similarly, the comparisons of the mean scores of actual and mediated scores of the 44 test takers showed a change of mean scores from 17.95 (SD=8.41) to 41.29 (SD=7.13).

The results of paired-sample t-test for students who took the difficult test also revealed that the difference between the actual and mediated scores was significant (t

Table 6 Descriptive and Inferential Statistics of Non-DA, DA, and Gain Scores on the Difficult Test

	N	Mean	SD	t (*p*)
Non-DA score	44	17.95	8.41	11.73 (0.000)
DA score	44	41.29	7.13	
Gain score	44	23.34	13.19	

Table 7 Descriptive Statistics of LPSs on the Easy and Difficult Tests

	N	Minimum	Maximum	Mean	SD
Easy test	31	0.56	1.62	0.98	0.26
Difficult test	44	0.23	1.41	0.82	0.24

(43) = 11.73, $p < .01$, $d = 2.99$). Furthermore, to reduce the likelihood of a Type I error, i.e., spuriously significant difference, the Bonferroni adjustment was conducted for each test. The desired alpha level (0.05) was divided by the number of comparisons made (i.e., 2 per test) and the Least Significant Difference (LSD) p-value required for significance would be 0.05/2 = 0.025. Since the p-value levels of the two comparisons were lower than the adjusted alpha level ($p = .00 < .025$), it can be concluded that the pairs of the actual and mediated scores for the easy and difficult tests showed significant differences.

It is worth mentioning that a highly significant correlation ($r = .52$, $p = .00$) was found between actual and mediated scores of the learners on the difficult test indicating a direct mapping across scores. The results also showed a strong negative correlation between gain scores and actual scores ($r = -.87$, $p = .00$), indicating that learners who did not perform well independently benefited more from mediation than those who performed better independently. This supports the positive effect of graduated hints on the development of learners' listening comprehension and is evidence of learners' internalization of mediation. With regard to the first research question, it can be concluded that the adaptive C-DA was able to improve the learners' listening ability.

4.2 Research question 2

The second research question explored whether the adaptive C-DA was able to reveal the learners' potential for learning. Here, the researchers aimed to address the ability of C-DA to assess the size of learners' ZPD. Therefore, Kozulin and Garb's (2002) formula was employed to calculate LPSs. Table 7 tabulates the mean LPSs of the learners for the two tests.

As depicted, the mean LPSs of the learners ranged from 0.56 to 1.62 on the easy test, and from 0.23 to 1.41 on the difficult test. As the results indicate, the students who made considerable progress from actual to mediated scores had high LPSs, and the learners who made slow progress had low LPSs.

To provide a more vivid picture of learners' LPSs, ten learners either with similar actual scores or exactly similar/similar ranges of LPSs were selected to check the degree of instructional help they needed. In so doing, the participants whose actual scores or LPS ranges were similar were divided into two subgroups. Then 10 participants were randomly selected from these 2 subgroups through stratified random sampling. Table 8 presents the actual, mediated, and LPSs of these ten learners

Table 8 The Non-DA, DA, and LPSs of the Learners on the Easy Test

Learners	NDA	DA	LPS
Reza	20	42	0.81
Simin	22	43	0.81
Zahra	8	26	0.56
Sarah	8	35	0.78
Nima	20	52	1.06
Ali	8	57	1.34
Tara	22	37	0.66
Elnaz	12	48	1.06
Melika	18	54	1.14
Soroush	26	45	0.81

on the easy test. For example, Reza and Nima had similar actual scores (20). Reza scored 42 under mediation, showing a mid-range LPS of 0.8, but Nima scored 52, demonstrating a high-range LPS of 1.06. Regarding this difference, it can be noted that although the two learners demonstrated similar independent performances, they needed different degrees of instructional help as they continued to develop their listening comprehension. Moreover, learners who received low actual scores (Zahra, Sarah & Ali) improved differently under mediation and hence gained different LPSs. With mediation Zahra and Sarah scored 26 and 35, showing low- and mid-range LPSs of 0.56 and 0.78, respectively. However, Ali scored 57 under mediation and gained a high-range LPS of 1.34.

It should be noted that actual scores demonstrate an already developed ability at the time of assessment and do not reveal learners' ZPD which is vital for diagnosis and future learning and teaching. Reporting actual and mediated scores, on the other hand, provides insight into a learner's incomplete and potential abilities. LPS completes this by quantifying the observed changes, the same as a gain score, but brings forward the results in relation to the maximum possible score. In this way, a learner with a low actual score is not harshly judged and may still be accepted to have a high-range LPS, as is the case with Ali shown in Table 8.

As it is shown, learners' LPSs on this test were not the same for learners with similar actual scores. Therefore, it can be concluded that the adaptive C-DA revealed L2 learners' potential for learning.

4.3 Research question 3

The third research question of the study examined whether the adaptive C-DA was capable of developing the internalization of mediation previously offered. Accordingly, the data obtained from transfer items were analyzed. Transfer items were included in each test without any overt indication of their function. The tests involved eight transfer items (4 per test). The transfer and non-transfer items were similar in terms of response options, hints and focus on the particular construct or item mode. They were only different in terms of difficulty level (transfer > non-transfer). Of course, the transfer items were not intended to simply determine whether learners could answer a particular grammatical or vocabulary item based on the listening test. Rather, they were intended to determine if the learners were better able to comprehend aural texts or if they continued to experience the same level of struggle with

Table 9 The Non-DA, DA, LPS, and Transfer Scores of the Learners on the Easy Test

Learners	NDA	DA	LPS	T
Reza	20	42	0.81	8
Simin	22	43	0.81	12
Zahra	8	26	0.56	8
Sarah	8	35	0.78	16
Nima	20	52	1.06	14
Ali	8	57	1.34	14
Tara	22	37	0.66	8
Elnaz	12	48	1.06	12
Melika	18	54	1.14	15
Soroush	26	45	0.81	16

Table 10 Mean and Standard Deviation of Transfer Scores Grouped by LPS Range on Easy Test

LPS range	*Mean (SD)*	*N*
High	16.75 (6.99)	15
Mid	15.86 (4.65)	13
Low	10 (5.35)	3

Table 11 Mean and Standard Deviation of Transfer Scores Grouped by LPS Range on Difficult Test

LPS range	*Mean (SD)*	*N*
High	17.36 (4.36)	11
Mid	15.41 (4.95)	29
Low	10 (5.35)	4

regard to different dimensions and item modes of the listening test. To better capture the picture, Table 9 demonstrates the actual, mediated, LPS, and transfer scores of the ten learners mentioned previously.

As it is shown in Table 9, Reza scored 8 in transfer items answering two comprehension-type items correctly without mediation. However, Nima, with the same actual score (20), scored 14 in transfer items answering three transfer (2 comprehension- & 1 production-type) items correctly without mediation. Moreover, a significant correlation was detected between transfer and LPSs ($r = .31$, $p = .03$ for the difficult test; $r = .79$, $p = .00$ for the easy test). A comparison of transfer scores on each LPS level was also performed. Table 10 presents the results.

As expected, learners with high LPS gained a higher mean (M = 16.75, SD = 6.99) for transfer scores followed by transfer scores of medium (M = 15.86, SD = 4.65) and low (M = 10, SD = 5.35) LPS levels.

Mann-Whitney U tests were run to check the significance of these differences in transfer scores. The Mann-Whitney U test was employed since we had non-parametric data. The result of Mann-Whitney U test revealed no significant differences in terms of transfer scores between the high and mid ($z = -0.44$, ns) and high and low ($z = -1.62$, ns) ranges of LPSs. However, there was a significant difference with regard to transfer scores between the mid and low ($z = -2.11$, $p = .03$) ranges of LPSs on the easy test.

The mean and standard deviation of transfer scores were also computed on the difficult test. Table 11 depicts the results.

As presented in the table, the mean and SD turned out to be 17.36 and 4.36 for the high LPS level, 15.41 and 4.95 for the mid LPS level, and 10 and 5.35 for the low LPS level, respectively. The results of the Mann-Whitney U test revealed no significant difference in terms of transfer scores between the high and mid (z= -1.17, ns), and mid and low (z= -1.52, ns) ranges. However, there was a significant difference between the high and low (z= -2.36, p=.01) ranges. Furthermore, the results of correlation and the comparison of mean scores indicated that LPS was promising in predicting learning. Therefore, it can be concluded that the adaptive C-DA had the potential of increasing the internalization of the mediation in the learners who took part in this study.

5 Discussion

This study explored the feasibility of an adaptive C-DA in terms of its potential in developing learners' listening comprehension, revealing learners' potential level of performance and their degree of internalization of mediation. Quantitative analyses were done and the findings revealed that the adaptive C-DA possessed these three different potentials.

Concerning the first research question, the results are in line with Poehner and Lantolf (2013) and Poehner et al. (2015). Poehner's studies with a focus on C-DA and the current study with a focus on adaptive C-DA revealed that DA could create a supportive atmosphere aiming at prioritizing learners' future learning and development by taking into account learners' ZAD and ZPD. While non-dynamic tests can only pay attention to intramental, self-regulated, and fully-internalized abilities of the learners, DA accounts not only for these abilities but also for the abilities which are other-regulated/intermental (Poehner, 2008). As Ahmadi and Barabadi (2014) pointed, "the significant gain of test takers from non-dynamic to dynamic test can be attributed to non-intellective factors … non-intellective factors such as lack of motivation, fear of failure, and inattentiveness can be the cause of incorrect response by test takers" (p. 176). By the same token, learners in the current study could answer correctly with the implicit hints and overcome these non-intellective factors that might have caused them to lose the whole score in a non-dynamic test. Furthermore, learners' significant gain in adaptive C-DA can be regarded as proof of the construct validity of DA. Construct validity, as Haywood and Lidz (2007) and Poehner (2008) stated, demonstrates the extent to which DA enhances individuals' development.

Regarding the second research question and adaptive C-DA's ability to measure learners' potential for learning, the obtained results support Kao and Kuo (2021) on listening comprehension and Yang and Qian (2017) on reading comprehension. They concluded that the deployment of C-DA could afford a vivid picture of EFL learners' independent and mediated performance and their listening and reading difficulties. In the same vein, Bakhoda and Shabani (2018) unraveled the capability of fine-tuned C-DA mediation to respond to individual learning preferences in assessing reading comprehension. However, the results obtained are not in line with Hidri and Pileh Roud (2020) who reported no empirical evidence for C-DA potential to provide a comprehensive diagnosis of the learners' ability in ZAD. In this regard, a discussion

of LPS and transfer scores seems necessary. According to Kozulin and Garb (2002), the ZPD level of learners who have high LPS is close to their ZAD denoting that "the targeted ability is on the verge of internalization or self-regulation" (Ahmadi & Barabadi, 2014, p. 176). On the other hand, the ZPD level of learners who have low LPS is not close to their ZAD denoting that the learners need much more mediation to internalize the given point. In line with Kozulin and Garb (2002), this study revealed that learners with low LPS used much more mediation in form of hints than learners with high LPS. This information concerning learners' learning potential scores can be used to create opportunities to develop their performance ability. For instance, learners who scored an average non-DA but a low-level LPS should do some exercises which would help them learn better in future activities. While those who scored an average non-DA but earned high LPS should do more challenging exercises. For the learners who scored poorly both in non-DA and LPS, some extra activities might be designed.

Moreover, taking the third research question into account, similar to Poehner's studies, this study revealed that LPS is capable of differentiating between learners with the same actual score. Poehner and Lantolf (2013) argued that actual scores do not directly reflect learners' abilities because identical actual scores do not inevitably map onto the same mediated scores. According to their study, of the three learners (6, 13, & 14) who produced the same actual score of 16 on the listening test, "learners 13 and 14 produced similar mediated scores, 29 and 28, respectively – while learner 6 produced a much higher mediated score, an indication that the learner responded more favorably to mediation" (p. 335). Importantly, while some learners with high actual scores did not improve as much under mediation and thus produced medium-level LPSs, some other learners with low actual scores did better under the mediation. Similarly, Yang and Qian (2020) maneuvered on transfer scores and reported on C-DA ability in supporting the students' potential for dealing with more difficult and more challenging tasks. Adversely, Ebadi and Rahimi (2019), in assessing the effect of C-DA mediations on learners' writing performance in new and more challenging contexts during near and delayed synchronous transcendence sessions, expressed learners' difficulties in transferring some parts of their developed writing skills to more challenging tasks. It can be discussed that actual scores are only able to uncover the abilities that have already matured and do not uncover the abilities that are developing or need to be developed. In this sense, traditional assessments miss the opportunities to enhance learners' abilities. It pinpoints that, as Alderson et al. (2015) argued, while the chief goal of any language assessment is to provide more detailed information for decision-making and diagnosis so that high and low stakeholders make use of them, DA in general and adaptive C-DA in particular lead to beneficial consequences for all parties involved.

Furthermore, a traditional view of assessment by including the same pre-planned set of items for all learners regards learners' difficulties as analogous (Race et al., 2005) and pays little attention to each learner's individual needs and underlying potentials. As a reaction to static assessment, DA has gained a foothold in assessing learners through teacher intervention aiming at expanding their learning abilities and at the same time striving to diagnose and assess learners' development. As a subset

of DA, C-DA adds to DA potentials and advantages through calibrated electronic interventions.

The effectiveness of C-DA can be discussed in terms of its distinct advantage in the simultaneous assessment and diagnosis of a large number of individuals. Moreover, technology has inaugurated computerized assessment as an invisible teacher capable of scoring, monitoring, and supervising the test while presenting focused mediational support to accommodate the learners' needs. The diagnosis provided through adaptive C-DA as a response to students' diversity demands knowledge of each learner's independent performance, mediated performance, responsiveness to mediation, as well as future performance through attempting a more challenging task. This focus on different aspects of each learner's current and future abilities aims at transcending the learner's current ability level not obtaining a higher score. Taking into account the learners' different needs and abilities, adaptive C-DA proved effective in identifying each student's strengths and needs in listening comprehension and providing each learner with instruction to address his/her individual learning needs. In fact, adaptive C-DA can significantly reduce the amount of time needed for assessment by eliminating and excluding those items too easy or too difficult for each individual. At the same time, it provides an efficient way for real-time feedback in line with each examinee's ongoing performance.

6 Conclusion

When it comes to language learning, computers and smart devices are ideal tools "to help raise the language awareness of students with varying needs and learning habits" (Dodigovic, 2003, p. 202). Adaptive C-DA programs with their instant access, analytical strength, tailored to different needs, and their ability to go through the same motions over and over again are unbeatable.

The adaptive C-DA with its focus on tailoring the difficulty of the test to the learners' abilities and adaptation of prompts to the examinees' needs can be innovative not only in the field of language testing but also in the field of DA. This adaptive tutor allows the simultaneous assessment of a large number of learners in a dynamic way and removes the necessity of the presence of instructors and learners in the class because computers and smart devices can assume the role of expert mediators. Moreover, the adaptive C-DA is capable of tracking learners' errors and can highlight learners' strong and weak points. Hence, instructors can switch the focus of instruction on learners' problematic areas. In this way, learners are able to self-assess and reassess themselves and get involved in the learning and assessment process. Accordingly, with the presence of adaptive C-DA learners "are no longer dependent upon teachers to be assessed and become aware of their progress; they can assess and reassess themselves as many times as needed" (Ahmadi & Barabadi, 2014, p. 180). Another advantage is that it can accommodate individual learners to a modest extent. While intelligence, language aptitude, and affective factors seem very crucial in an individual's learning prosperity, the adaptive C-DA has not yet attempted to support individual differences at this level. It should be mentioned that this program is like a human baby and has a lot more to learn and accomplish. In other words, like a human

baby, the program has the ability to acquire "more and more complex 'knowledge' that will eventually bring about maturity" (Dodigovic, 2003, p. 202).

Generally speaking, the results of this study lay a path between technology and assessing learners' performance and establish a solid basis for the integration of intelligent computer programs within teaching and assessment tasks and using them to improve one's performance. Some steps have been undertaken in the current attempt that needs backing from institutions' principals, educators, and curriculum developers for a significant leap forward. The assessment tool proposed here is teacher- and learner-friendly and it can be easily incorporated into any instructional setting with learners from different age groups.

However, there were some limitations to be acknowledged and addressed regarding the present study. The adaptive C-DA devised in this study is an attempt in developing an assisted language assessment-learning device that includes natural human processing in the form of graduated hints. The prompts were developed based on a meticulous analysis of mediator-learner interaction. Anyhow, it was the first time that learner reciprocity was considered in developing prompts. The devised C-DA was tested and retested before its final use; nevertheless, it is not without shortcomings that should be addressed in future studies. It is worth mentioning that the shortcomings do not seem to pose a serious threat to the role of natural language processing in adaptive C-DA and raising language awareness. But overcoming these shortcomings leads to further improvement in C-DA and makes it available to a broad public in a more user-friendly form.

To expand the use of adaptive C-DA, future studies can examine the correlation between the learners' choices when working with the program and their learning styles. For example, it can be investigated whether analytical learners really like to accomplish the task by themselves rather than being provided with the correct response. A comparative study based on gender in terms of C-DA effectiveness in assessing and diagnosing listening comprehension or other language skills provides another promising direction for future research. Another venue for future studies is the use of the adaptive C-DA for a longer period of time. In the present study, the program supported a very small curriculum, which would not stretch over a longer period of time. Therefore, a program applied over a longer period of time would be worthy of examination.

Authors' contributions Mehri Izadi developed the listening comprehension tests and gathered and analyzed the data.
Maliheh Izadi designed and developed the adaptive Computerized Dynamic Assessment Tutor.
Farrokhlagha Heidari interpreted the statistical analyses, wrote the manuscript, and revised and carried out the corrections as suggested by reviewers.
All authors read and approved the final manuscript.

Funding No funding was received for conducting this study.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

References

Ableeva, R. (2010). *Dynamic assessment of listening comprehension in Second Language Learning* [Unpublished doctoral dissertation]. Pennsylvania State University. https://etda.libraries.psu.edu/catalog/11063.

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

Ahmadi Safa, M., & Beheshti, S. (2018). Interactionist and interventionist group dynamic assessment (GDA) and EFL learners' listening comprehension development. *Iranian Journal of Language Teaching Research*, *6*(3), 37–56. https://doi.org/10.30466/IJLTR.2018.120600.

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics*, *36*(2), 236–260. https://doi.org/10.1093/applin/amt046.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Bakhoda, I., & Shabani, K. (2018). Bringing L2 learners' learning preferences in the mediating process through computerized dynamic assessment. *Computer Assisted Language Learning*, *32*(3), 210–236. https://doi.org/10.1080/09588221.2018.1483950.

Barabadi, E., Khajavy, G. H., & Mehri Kamrood, A. (2018). Applying interventionist and interactionist approaches to dynamic assessment for L2 listening comprehension. *International Journal of Instruction*, *11*(3), 681–700. https://doi.org/10.12973/iji.2018.11346a.

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. https://doi.org/10.1016/j.sbspro.2012.01.006.

Burston, J., Neophytou, M., & Lamprianou, I. (2016). Developing and implementing a computer-adaptive test for English: The SLUPE experience. *CALL-EJ*, *17*(1), 19–34.

Chen, J., & Wang, L. (2010, October 29-30). Computerized Adaptive Testing: A New Trend in Language Testing. In *International Conference on Artificial Intelligence and Education (*ICAIE) [Paper presentation]. IEEE. https://doi.org/10.1109/ICAIE17076.2010

Dodigovic, M. (2003). Natural Language Processing (NLP) as an instrument of raising the language awareness of learners of English as a second language. *Language Awareness*, *12*(3–4), 187–203. https://doi.org/10.1080/09658410308667076.

Dörfler, T., Golke, S., & Artelt, C. (2017). Evaluating prerequisites for the development of a dynamic test of reading competence: Feedback effects on reading comprehension in children. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education. Methodology of educational measurement and assessment* (pp. 487–503). Springer.

Ebadi, S., & Rahimi, M. (2019). Mediating EFL learners' academic writing skills in online dynamic assessment using Google Docs. *Computer Assisted Language Learning*, *32*(5–6), 1–29. https://doi.org/10.1080/09588221.2018.1527362.

Ebadi, S., & Saeedian, A. (2019). Exploring L2 learning potential through computerized dynamic assessment. *Teaching English Language*, *13*(2), 51–78. https://doi.org/10.22132/TEL.2019.92190.

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Green, A. B. (2016). Assessment literacy for language teachers. In D. Tsagari (Ed.), *Classroom-based assessment in L2 contexts* (pp. 8–29). Cambridge Scholars.

Guzmán, E., & Conejo, R. (2005). Self-assessment in a feasible, adaptive web-based testing system. *IEEE Transactions on Education*, *48*(4), 688–695.

Haerazo, J. D., Davin, K. J., & Sagre, A. (2019). L2 dynamic assessment: An activity theory perspective. *The Modern Language Journal*, *103*(2), 443–458. https://doi.org/10.1111/modl.12559.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. https://doi.org/10.1177/0265532214564505.

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

Heidari, F., & Izadi, M. (2020). The potentiality of dynamic assessment in Massive Open Online Courses (MOOCs): The case of listening comprehension MOOCs. *Iranian Journal of English for Academic Purposes*, *9*(2), 1–13.

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

Kao, Y. T., & Kuo, H. C. (2021). Diagnosing L2 English learners' listening difficulties and learning needs through computerized dynamic assessment. *Interactive Learning Environments*. Advance online publication. https://doi.org/10.1080/10494820.2021.1876738.

Khoshsima, H., & Izadi, M. (2014). Dynamic vs. standard assessment to evaluate EFL learners' listening comprehension. *Iranian Journal of Applied Language Studies*, *6*(2), 1–26. https://doi.org/10.22111/IJALS.2014.2187/.

Kozulin, A., & Garb, E. (2002). Dynamic assessment of EFL text comprehension. *School Psychology International*, *23*(1), 112–127. https://doi.org/10.1177/0143034302023001733.

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics*, *1*(1), 49–72. https://doi.org/10.1558/japl.1.1.49.55872.

May, L., Nakatsuhara, F., Lam, D., & Galaczi, E. (2019). Developing tools for learning-oriented assessment of interactional competence: Bridging theory and practice. *Language Testing, 37*(2), 165–188. https://doi.org/10.1177/0265532219879044.

Mehri Kamrood, A., Davoudi, M., Ghaniabadi, S., & Amirian, S. M. R. (2021). Diagnosing L2 learners' development through online computerized dynamic assessment. *Computer Assisted Language Learning*, *34*(7), 868–897. https://doi.org/10.1080/09588221.2019.1645181.

Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the word part levels test. *Language Testing, 36*(1), 101–123. https://doi.org/10.1177/0265532217725776.

Oppl, S., Reisnger, F., Ekmaier, A., & Helm, C. (2017). A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education*, *14*(2), 1–21. https://doi.org/10.1186/s41239-017-0039-0.

Pishghadam, R., Barabadi, E., & Mehri Kamrood, A. (2011). The differing effect of computerized dynamic assessment of L2 reading comprehension on high and low achievers. *Journal of Language Teaching and Research*, *2*(6), 1353–1358. https://doi.org/10.4304/jltr.2.6.1353-1358.

Poehner, M. E. (2005). *Dynamic assessment of oral proficiency among advanced L2 learners of French* [Unpublished doctoral dissertation]. Pennsylvania State University, University Park, PA. https://etda.libraries.psu.edu/catalog/6627

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

Poehner, M. E., & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research*, *9*(3), 233–265. https://doi.org/10.1191/1362168805lr166oa.

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. https://doi.org/10.1177/1362168813482935.

Poehner, M. E., Zhang, J., & Lu, X. (2015). Computerized Dynamic Assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing, 32*(3), 337–357. https://doi.org/10.1177/0265532214560390.

Race, P., Brown, S., & Smith, B. (2005). *500 Tips on assessment* (2nd edition.). Routledge.

Shabani, K. (2012). Dynamic assessment of L2 learners' reading comprehension processes: A Vygotskian perspective. *Procedia-Social and Behavioral Sciences, 32*, 321–328. https://doi.org/10.1016/j.sbspro.2012.01.047.

Shabani, K. (2014). Dynamic assessment of L2 listening comprehension in transcendence tasks. *Procedia-Social and Behavioral Sciences*, *98*, 1729–1737. https://doi.org/10.1016/j.sbspro.2014.03.600.

Shrestha, P., & Coffin, C. (2012). Dynamic assessment, tutor mediation, and academic writing development. *Assessing Writing*, *17*, 55–70. https://doi.org/10.1016/j.asw.2011.11.003.

Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education*, *97*, 69–85. https://doi.org/10.1016/j.compedu.2016.02.018.

Van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. Springer Publishing.

Veldkamp, B. P., & Matteucci, M. (2013). Bayesian computerized adaptive testing. Ensaio: Avaliação e Políticas Públicas em Educação, *21*(78),57–82. https://doi.org/10.1590/S0104-40362013005000001

Vygotsky, L. S. (1998). The problem of age. In R. W. Rieber (Ed.), *Child psychology: Collected works of L.S. Vygotsky* (pp. 187–205). Plenum.

Wainer, H. (2000). *Computerized adaptive testing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, T. (2010). Web-based dynamic assessment: Taking assessment as teaching and learning strategy for improving students' e-learning effectiveness. *Computers & Education*, *54*(4), 1157–1166. https://doi.org/10.1016/j.compedu.2009.11.001.

Ward, W. C., & Bennett, R. E. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (1st ed.). Routledge. https://doi.org/10.4324/9780203052518.

Yang, Y., & Qian, D. D. (2017). Assessing English reading comprehension by chinese EFL learners in computerized dynamic assessment. *Language Testing in Asia*, *7*(11), 1–15. https://doi.org/10.1186/s40468-017-0042-3.

Yang, Y., & Qian, D. D. (2020). Promoting L2 English learners' reading proficiency through computerized dynamic assessment. *Computer Assisted Language Learning*, *33*(5–6), 628–652. https://doi.org/10.1080/09588221.2019.1585882.

Zangoei, A., Zareian, G., Adel, S. A. M., & Amirian, S. A. M. (2019). The impact of computerized dynamic assessment on Iranian EFL learners' interlanguage pragmatic development. *Journal of Modern Research in English Language Studies, 6*(4), 139–165. https://doi.org/10.30479/jmrels.2019.11536.1433.

Authors and Affiliations

Mehri Izadi[1] · Maliheh Izadi[2] · Farrokhlagha Heidari[1]

✉ Farrokhlagha Heidari
heidari.f@english.usb.ac.ir

Mehri Izadi
izadimi@yahoo.com

Maliheh Izadi
m.izadi@tudelft.nl

[1] Department of English Language and Literature, University of Sistan and Baluchestan, Zahedan, Iran

[2] Software Engineering Research Group, Delft University of Technology, Delft, The Netherlands