Language Testing in Asia
a SpringerOpen Journal

RESEARCH Open Access

Developing and evaluating a dynamic assessment of listening comprehension in an EFL context

Sahbi Hidri

Correspondence:
sahbihidri@gmail.com
Faculty of Human and Social Sciences of Tunis, 94, BD du 9 avril 1938, Tunis 1007, Tunisia

Abstract

This study addressed a need to examine and improve current assessments of listening comprehension (LC) of university EFL learners. These assessments adopted a traditional approach where test-takers listened to an audio recording of a spoken interaction and then independently responded to a set of questions. This static approach to assessment is at odds with the way teaching listening was carried out in the classroom, where LC tasks often involved some scaffolding. To address this limitation, a dynamic assessment (DA) of a listening test was proposed and investigated. DA involves mediation and meaning negotiation when responding to LC tasks and items. This paper described: (a) the local assessment context, (b) the relevance of DA in this context, and (c) the findings of an empirical study that examined the new and current LC assessments. Sixty Tunisian EFL students responded to a LC test with two parts, static and dynamic. The tests were scored by 11 raters. Both the test-takers and raters were interviewed about their views of the two assessments. Score analyses, using the Multi-Facet Rasch Measurement (MFRM) (FACETS program, version, 3.61.0), indicated that test-taker ability, rater behavior and item difficulty estimates varied across test types. Qualitative data analysis indicated that although the new assessment provided better insights into learners' cognitive and meta-cognitive processes than did the traditional assessment, raters were doubtful about the value of and processes involved in DA mainly because they were unfamiliar with it. The paper discussed the findings and their implications for listening assessment practices in this context and for theory and research on listening assessment.

Keywords: Dynamic/Static assessment; Ability estimates; Rater behavior; Item difficulty; Significant bias; FACETS; Qualitative; Quantitative analysis

Background

The purpose of this study addressed a need to examine and improve current assessments of listening at the tertiary level. In this study, two listening tests, dynamic and static, were examined and assessed. Static LC tests have been used in language research and assessment. This type of listening seems to be at odds with the way teaching listening is carried out in class in which learners are supposed to be engaged in joint activities to comprehend listening. Static assessment (SA) rests on engaging the test-takers in working on the test individually with no scaffolding on the part of mediators or test-takers. SA may be more convenient and practical than DA, especially in large-scale situations (Lantolf & Poehner 2010). In static or traditional LC tests, there is no interest allocated

Springer

© 2014 Hidri; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

to the joint interactions of the learners required for approaching the learning input (Leung 2007; Lidz & Gindis 2003).

The pendulum in language testing research has shifted to the social dimension of language testing where learners are tested on their abilities to use language in a particular social setting (McNamara 2001; McNamara & Roever 2006). For instance, a growing interest has been given to the link between Second Language Acquisition (SLA) and language assessment (Bachman & Cohen 1998; Chalhoub-Deville 2003; Douglas & Selinker 1985; Lantolf 2009; Leung 2007). McNamara (2000) highlights this emerging trend in testing when he states that

> New forms of language assessment may no longer involve the ordeal of a single test performance under time constraints. Learners may be required to build up a portfolio of written or recorded oral performances for assessment. They may be observed in their normal activities of communication in the language classroom on routine pedagogical tasks. […] Pairs of learners may be asked to take part in role plays or in group discussions as part of oral assessment. (p. 4)

Many researchers (e.g., Lantolf & Poehner 2006; Ohta 2000; Swain 2000) argue that language acquisition and learning can be achieved through joint interactions. Such an interaction can be implemented through using prompts, hints, clarifications, and leading questions. In part, the use of these strategies depends on the language ability of the learner. Chalhoub-Deville (2003), p. 377 claims that "it is likely that language users at different proficiency levels call upon different or differentially developed abilities." Since dynamic listening tasks involve interaction among students and guided performance of learners by mediators (Gibbons 2003; Lantolf & Poehner 2004, 2006), it is no wonder then that such tests, necessitating interaction, can inform about language learning and assessment.

DA can be traced back to Vygotsky (1981, 1986) that stresses the social environment as a facilitator of the learning process (Karpov & Haywood 1998; Kozulin & Garb 2002). DA has gained momentum in research (e.g., Leung 2007; Poehner and van Compernolle 2011; Rea-Dickins 2006; Tzuriel 2011) and has also been applied to classroom-based assessment (Ableeva 2008; Ableeva & Lantolf 2011; Sternberg & Grigorenko 2002). In DA, teaching and testing are intertwined into a joint activity which targets the activation of the learners' cognitive and metacognitive processes (Ableeva & Lantolf 2011; Tzuriel 2011, p.115). Research (e.g., Gass 1997; Lidz 2002; Swain 2001) has shown that learners become co-constructors of meaning in collective joint activities where knowledge and meaning can be negotiated and mediated. This negotiation is context-bound.

Mediation, the zone of proximal development (ZPD), contingency and scaffolding are cornerstones in DA. Vygotsky's theory of learning stresses mediation in that it can instruct learners in how to use their cognitive and metacognitive strategies, for instance, in a problem-solving activity. Gibbons (2003), p. 249 defines the ZPD as "the cognitive gap between what learners can do unaided and what they can do in collaboration with a more competent other". To this end, learners can only perform successfully in the presence of another participant, such as a teacher. Contingency consists of the "assistance required by the learner on the basis of moment-to-moment understanding" (Gibbons 2003, p. 267) i.e., teachers modulates the kind of support based on the learners'

reaction of and attitudes towards this support. Scaffolding, however, mediates the learners in acquiring new strategies to be able to finish the task independently (Kozulin & Garb 2002). This requires activation of the cognitive and meta-cognitive strategies to be able to comprehend the listening input. An awareness of such strategies can be conducive to success in language learning and assessment. In this regard, Vandergrift et al. (2006) note that awareness of the listening strategies "can have positive influence on language learners' listening development" (p. 432) and by extension to accessing the test items easily. Such awareness is a cornerstone in assessing LC dynamically. Adhering to DA both in teaching and testing depends on the teaching experience, experience with language, motivation and views of language and language learning.

The duality between dynamic and static assessment can in fact be blended together with the goal of forming a comprehensive view about the LC ability of the test-takers. Though complementary they might appear, static and dynamic assessment have methodological differences. Since this type of assessment considers the learners' abilities as already matured i.e., fixed and "stable across time" (Leung 2007, p. 260), in DA, such abilities are "malleable and flexible" (Sternberg & Grigorenko, 2002, p. 1). In addition, while scores in SA may be praised for their objectivity, they nevertheless fail to infer much about the learners' cognitive processes. Hence, the importance of implementing DA. SA focuses on the product of learning; however, in DA, much importance is given to developing learning in that the main focus is attributed to the processes which lead to the end product. Proponents of DA highlight the idea that such an assessment mode should not lead to failure; rather it should be conducive to better linguistic attainment. Most studies on DA (e.g., Ableeva 2008; Gibbons 2003; Lidz 2001) have shown that after mediation takes place, learners can reach higher levels of much scaffolding. Because of its receptive nature, listening test items should be processed dynamically.

Different studies have been carried out to address traditional or static LC from different angles, such as the effects of background knowledge on listening performance (Jensen & Hansen 1995), the use of LC cognitive processes to comprehend the listening input (Buck & Tatsuoka 1998), effects of speech rate on item difficulty (Brindley & Slayter 2002), the use of multiple-choice (MC) format and its impact on test scores (Yi'an 1998). Ginther (2002) investigated the effects of content visuals on LC in the TOEFL test in LC passages of different genres, such as dialogues, short conversations, academic discussions and mini-talks. Also, Berne (1995) addressed the variation of pre-listening activities and its impact on LC; while Rubin (1994) dealt with the effects of top-down and bottom-up processes on comprehension of listening. However, compared to static listening, few studies have investigated dynamic listening. For instance, Ableeva (2008) addressed the effects of DA on comprehending listening. Ableeva and Lantolf (2011), in a longitudinal qualitative study, highlighted the importance of using DA in developing the mental processes of comprehending listening in French. In addition, while research on testing has been concerned with the joint interactions in language skills such as speaking (Fulcher 1996; Swain 2001), scant attention has been allocated to such interactions between teachers and students in other language skills, such as listening. One way of approaching alternative assessments to LC in interactive patterns can be carried out through DA. To date, approaching both modes of assessment in LC, i.e., static and dynamic, has received scant attention in language testing research.

Assessment context and relevance of DA

In testing LC in Tunisia, test-takers have always been given an audio-taped passage to listen to and then respond to a limited set of test items, such as wh-questions and true/false statements; thus underrepresenting the LC construct which was supposed to embrace as many LC test items as possible (Hidri 2010a, 2013a, 2013b, 2014). By limiting testing to a very narrow range of skills, test designers may miss the target of measurement. In the Tunisian context, testing has been marginalized in targeting a fair measurement that would reflect the actual language ability of the learner (Hidri 2010b, 2014). This marginalization is even echoed in teaching given the eventuality of the teacher being the resource.

In effect, the view of language learning consists of teachers doing most of the talking in class. For instance, Helal (1997) addressed the use of the communicative competence framework developed by Canale and Swain (1980) to the treatment of EFL learners' errors in Tunisia. Based on a questionnaire administered to teachers and cross-sectional visits to some EFL classes in Tunisia, Helal found out that most of the teachers' attention during classroom interaction was geared towards the treatment of students' grammatical errors even in tasks calling for greater attention to communication, discourse and sociolinguistic appropriateness. Especially relevant to this study is his conclusion that this is not surprising in Tunisian classes given the fact that learners are studying to pass exams which are still informed by structuralist and behavioral views of language and language learning. This view of teaching of most of Tunisian teachers is also reflected in testing. Further, there is an urgent call to reconsider and investigate the assessment practices in Tunisia in that graduates and post-graduates of English who embark on teaching at the vocational, primary, secondary and tertiary levels are not offered any course in testing as part of the curriculum. They are not even trained in how to carry out classroom-based assessment such as DA, nor are they exposed to developing effective teaching strategies that use scaffolding to help learners overcome their listening difficulties. This is the current situation now. They learn test design out of teaching experience.

There are three basic national exams at the primary, basic and secondary levels. According to officials in the Ministry of Education and some ELT inspectors at the secondary level in Tunisia, more than 77% of the Baccalaureate[a] students got below the score of 9.99 out of 20 in the English exam for the year 2012[b]. Despite the fact that there is no testing course, the assessment policy at the tertiary level calls for administering 3 tests in all disciplines (2 progress and 1 achievement tests) over a fourteen-week term. Students rarely study 14 weeks and they most often tend to be absent from class even though there is a compulsory attendance policy. Because of these tremendous difficulties these learners have been facing in English, employing DA in a skill like listening may help such learners overcome these learning difficulties or change their learning behaviour.

The purpose of this study was motivated by three basic issues the first of which was the need to investigate the traditional and current assessment practices to determine the key idea that DA is meant to promote the test-takers' mental processing and their capacity to learn. Second, despite the fact that listening holds a major importance in learning and acquiring language, it has not been largely addressed in research compared to reading, writing, and speaking (Alderson 2005). Finally, research on

testing has only been concerned with static listening. Perhaps, the lack of concern for testing dynamic listening has been due to some practical issues, such as the difficulty of testing dynamic listening in large-scale situations on the one hand, and the scoring of the joint performance, on the other. There is a significant need to address these shortcomings. It could be then stated that the test-takers' listening ability in such contexts would vary from one test type to the other and that even the raters themselves would behave differently. It is in this context that it is crucial to investigate how these three variables of rater, test-takers' ability and item difficulty impacted static and dynamic listening assessment. Therefore, the study addressed the following research questions:

a. To what extent do estimates of test-taker ability and item difficulty vary across static and dynamic listening?
b. Is there any bias interaction of rater by test type, rater by test-taker and test-taker by test type?
c. What are the mediators' and test-takers' perceptions of both modes of assessment?

Methods

Participants

The 60 participants, who were selected to take part in this study, were first-year students majoring in English from a university in Tunisia. Previously, they had studied English as a required subject at school for seven years and were tested on listening at least twice a year. However, in the Baccalaureate exam, they were only tested on reading and writing. Students were admitted to university without any placement or diagnostic test. Before 2007, these students had to study four years to obtain their BA in English and they were supposed to teach English in secondary schools; while others who excelled were selected to sit for the MA program.

During the two first years of the curriculum at university, these participants had an oral skills course that combined both listening and speaking. They also took four listening exams, one in each term. Starting from 2006–2007, policy-makers initiated an ad hoc change of the educational system in Tunisia, by reducing the university study years from 4 to 3. This change also concerned courses and even the assessment system. For instance, the first-year participants, in the *licence*, *master*, *doctorat* (LMD[c]) system, were taught listening and speaking in two separate courses and were also supposed to study for three years, instead of four, to get their *licence*. All the participants were speakers of Tunisian Arabic, Modern Standard Arabic (MSA), French and English and they ranged in age from 19 to 21, with 47 females and 13 males.

All the 11 raters who took part in this study and who did both the mediating and rating were involved in teaching as well as testing LC. They all had an MA degree in applied linguistics, literature or culture studies and an English teaching experience that ranged from 1 to 14 years. For the sake of standardizing the scoring criteria, they all were engaged in training sessions in how to carry-out classroom-based assessment, such as role-plays and group discussions in dynamic listening. Then, the

researcher met with these teachers in their regular classes and were observed in how to carry out DA. After the course was over, the researcher evaluated the practicality of DA for improvement purposes. This had the purpose of helping the mediators become familiar with DA.

Data collection

There was a selection of a group composed of 60 test-takers. Test administration of the dynamic part was carried out during regular class hours as a progress test; while the static test was administered as a final achievement test, i.e., after one year of studying listening. The progress dynamic test was divided into three testing phases which were supposed to be dealt with in 45 minutes. It included 14 test items which were meant to generate negotiation of meaning between two test-takers and two mediators who also did the rating. The mediators offered support and guidance only in the pre- and while-testing phases. However, they were instructed to reduce mediation in the post-testing phase. The pre-testing phase, which lasted ten minutes included wh-, guessing and matching items. The while-testing phase lasted 20 minutes and it contained two wh- and summarizing items each, MC, true/false and guessing items. The raters were instructed to mediate the test-takers in both phases. The post testing phase lasted 15 minutes and it included MC, picture reordering, summarizing and making inference items.

In the one-hour achievement static test, the test-takers performed individually. This test included 40 items (five tasks with eight items each: Gap-filling, MC, information transfer, true/false statements and following instructions. This test was scored by 11 raters. The scores were analysed using the FACETS to account for ability estimates and item difficulty. Both participants were administered an interview to probe into their perceptions of and attitudes towards both parts of the test, particularly their degree of agreement with the practicality of the dynamic test, procedures of implementing, organizing the turn-taking of the joint interactions. In the static test, the raters were referred to as "raters", while in the dynamic test, they were referred to as "mediators." Table 1 reports the research design.

Methods of analysis

This study addressed a need to examine current assessments of LC of university learners of English. To address this, the quantitative and qualitative analyses were undertaken. Scores in the dynamic test were assigned once students finished providing their final answers. The quantitative analysis, relied on the use of the *FACETS* program (version 3.61.0) (Bond & Fox 2007) to analyse the scores of both parts of the test. Analysing test scores using the FACETS was used in research (e.g., Lumely & McNamara 1995; McNamara 1996, Kondo-Brown 2002). FACETS provides estimates of test-taker ability, item difficulty as well as biased interactions between elements of the different facets (e.g., rater by test type, rater by test-taker and test-taker by test type). Interview data were examined to identify patterns and themes in mediators' and test-takers' responses in relation to their perceptions of the dynamic test, its qualities, and feasibility of its use in a classroom-based assessment context.

Table 1 Research design

Research questions	Participants	Data	Analysis
a. To what extent do estimates of test-taker ability and item difficulty vary across static and dynamic listening?	60 test-takers who sat for both tests: Static and dynamic	***a.*** *A one-hour achievement static LC test*	*Quantitative analysis*Analysis of test scores:
b. Is there any bias interaction of rater by test type, rater by test-taker and test-taker by test type?	11 raters who scored both tests (in the static test, raters are referred to as "raters"; while they are referred to as "mediators" in the dynamic test.	Forty items (5 tasks with 8 items each) Item type: Gap filling, MC, information transfer, true/false statements and following instructions	- *FACETS*: Ability estimates and item difficulty
c. What are the mediators' and test-takers' perceptions of both modes of assessment?		***b.*** *A forty-five minute progress dynamic test*	- *Bias analyses of* rater by test type, rater by test-taker and test-taker by test type
		Fourteen items *Pre-testing phase* (10 minutes): Wh, guessing, matching	
		While-testing phase (20 minutes): Wh (x2), summarizing (x2), MC, true/ false, guessing	*Qualitative analysis*
		Post-testing phase (15 minutes): MC, picture reordering, summarizing, making inference	Analysis of the interview data
		c. *Interview* with raters and test-takers in both test modes	- Interview analysis: Perceptions of and attitudes towards the dynamic test

Results and discussion

The first part of this section addresses the FACETS analyses of test-taker ability and item difficulty reports. The second part reports the bias interaction of a) rater by test type, b) rater by test-taker and c) test-taker by test type. All these patterns were compared in both parts of the test to account for the sources of variability among the different facets.

Test-taker ability and item difficulty

To probe into the nature of the test-taker ability and item difficulty, the following question was addressed:

a. To what extent do estimates of test-taker ability and item difficulty vary across static and dynamic listening?

Table 2 describes the test-taker ability in both parts of the test. The ability logit value of the candidates in the dynamic part ranged from 3.19 to -.21 with candidate 37 being the most able (3.19 logits) and candidate 15 the least able (.-21 logits). The ability estimate mean for all the test-takers was 1.61. However, the ability values of the static test showed that there were less able test-takers, ranging from 2.48 to -.49

Table 2 Test-taker measurement report in both parts of the test

Dynamic				Static			
Test-taker	**Ability**	**SE**	**Infit MnSq**	**Test-taker**	**Ability**	**SE**	**Infit MnSq**
37	3.19	.76	1.00	23	2.48	.40	1.45
27	3.13	.76	1.02	22	2.41	.38	1.17
57	2.66	.64	.93	2	1.60	.30	1.77
5621	2.44	.57	.87	37	1.27	.27	1.39
	1.55	.49	.77	29	.40	.24	1.85
60	1.24	.48	1.63	56	1.12	.26	.95
23	1.22	.48	2.21	60	1.04	.26	.78
6	1.09	.47	1.96	57	.73	.25	.64
16	1.07	.47	1.08	18	.26	.24	.84
20	.88	.47	2.40	16	-.48	.24	.61
15	-.21	.46	1.68	19	-.49	.25	.67
N = 60				N = 60			
Mean	1.61	.53	1.00	Mean	.78	.26	1.03
SD	.82	.07	.40	SD	.66	.04	.31
Notes: Reliability of separation index = .53; fixed (all same) chi-square: 137.8 d.f.: 59 significance: p < .00				Notes: Reliability of separation index = .84; fixed (all same) chi-square: 308.5 d.f.: 59 significance: p < .00			

with an ability mean of .78. The infit mean square was 1.00 with a SD of .40 and a mean of 1.03 with a SD of .31 in the dynamic and static tests respectively. The consistency value, according to McNamara (1996), can be set using the mean with 2 SD in both directions. For instance, SD was .40 and the mean was 1.00 (.40×2 = .80 + 1.00 = 1.8), candidates 23, 6 and 20 in the dynamic and candidates 2 and 29 in the static parts were identified as misfitting. Misfitting candidates in the dynamic test might mean that the test mediators offered much supportive comments which led the candidates to perform better on the test items. However, the misfitting candidates in the static test could be due to, as McNamara (1996), p. 177 pointed out, "failure of attention in the test-taking process, guessing, anxiety, poor test item construction and the like." The reliability of separation index in the dynamic test was .53 and the chi-square of 137.8 with 59 d.f. was significant at $p < .00$. Also, the reliability of the separation index in the static test of the test was .84 and the chi-square of 308.5 with 59 d.f. was significant at $p < .00$. Therefore, the candidate ability estimates differed significantly in both test modes.

Table 3 shows four statistics: Item number, item difficulty, standard error (SE) and infit mean square of both tests. The difficulty mean value of the dynamic items was .00, ranging from .33 to -.28. The SE mean was .25, ranging from .24 to .26 with a SD of .01. In observing the static test, the difficulty mean value was .00 with a SD of .42, contrary to the dynamic test, ranging from .77 to −1.71. The SE mean was .21, ranging from .20 to 31. The most harshly scored item was item 108 (summarizing) with a logit difficulty of .33. The most leniently scored item was item 103 (matching) with a logit difficulty of -.28. The difficulty span of these two items was .61 (.33 + −.28). In the static test, the most harshly scored item was item 238 (following instructions), with a difficulty value of .77. The most leniently scored item was item 205 (gap filling), with a difficulty value of −1.71. The difficulty span was larger than the one of the dynamic

Table 3 Item measurement report in both parts of the test

Dynamic				Static			
Item	**Difficulty**	**Error**	**Infit MnSq**	**Item**	**Difficulty**	**Error**	**Infit MnSq**
108	.33	.24	.95	238	.77	.20	1.23
106	.28	.24	.87	218	.50	.20	1.29
101	.22	.24	.90	237	.46	.20	1.31
111	.16	.24	1.17	222	.26	.20	.64
114	.10	.25	.80	212	.01	.21	1.37
109	-.02	.25	1.06	216	-.08	.21	.53
110	-.02	.25	1.01	202	-.26	.22	1.31
113	-.02	.25	.95	206	-.31	.22	.86
112	-.08	.25	1.13	210	-.31	.22	.63
102	-.15	.25	1.24	207	-.40	.22	1.04
104	-.15	.25	1.10	239	-.40	.22	.98
107	-.15	.25	.90	221	-.60	.23	1.01
105	-.21	.26	.99	209	-.66	.23	1.01
103	-.28	.26	.96	205	−1.71	.31	1.21
N = 14				N = 40			
Mean	.00	.25	1.00	Mean	.00	.21	1.01
SD	.18	.01	.12	SD	.42	.02	.21
Notes: Reliability of separation index = .00; fixed (all same) chi-square: 7.7 d.f.: 13 significance: $p < .96$				*Notes*: Reliability of separation index = .74; fixed (all same) chi-square: 121.2 d.f.: 39 significance: $p < .00$			

test. The infit mean square was 1.00, ranging from .80 to 1.24; while it was 1.01 with a SD of .21 in the static test. This suggested that no item was identified as misfitting. The reliability of the separation index of the dynamic part of the test was very low .00 and the chi-square of 7.7 with 13 d.f. was significant at $p < .96$. Therefore, the null hypothesis that all the items were not equally difficult had to be rejected. In other words, the items did not differ significantly in terms of difficulty. The reliability of the separation index of the static part was .74 and the chi-square of 121.2 with 39 d.f. was significant at $p < .00$. Therefore, the null hypothesis that all the items were difficult had to be retained and confirmed.

Bias analyses

This section addressed the bias analyses of the interactions of rater by test type, rater by test-taker and test-taker by test type. It basically tried to answer the following question:

b. Is there any bias interaction of rater by test type, rater by test-taker and test-taker by test type?

Table 4 presents the bias analyses of the interaction of rater by test type. The data, sorted out according to the t value, column 9, revealed that there were 22 biased interactions out of the total count of measurable responses of 3240. Recall that both tests, dynamic and static, contained 14 and 40 items respectively (14 + 40 = 54x60 = 3240). Column 1 is the rater ID, column 2 is severity. The next four columns 3, 4, 5 and 6, show the total score

Table 4 Bias analysis of rater and test type

Rater	Severity	Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias	Error	t	Infit MnSq	Test type
5	-.36	119	109.8	70	.13	-.47	.24	−1.94	.9	Dynamic
6	.67	333	314.4	240	.08	-.20	.11	−1.91	.9	Static
11	-.42	73	70.8	42	.05	-.22	-.33	-.67	.9	Dynamic
10	-.18	66	64.1	42	.05	-.14	.28	-.51	1.0	Dynamic
3	.10	237	234.0	140	.02	-.08	.17	-.51	1.3	Dynamic
7	-.51	95	93.1	56	.03	-.14	.27	-.51	1.1	Dynamic
4	-.13	129	126.3	84	.03	-.10	.19	-.51	.9	Dynamic
1	.22	286	283.1	200	.01	-.04	.12	-.34	1.2	Static
2	.43	149	147.4	126	.01	-.03	.14	-.22	.9	Dynamic
8	-.03	147	146.5	120	.00	-.01	.14	-.08	.9	Static
9	.20	77	76.8	56	.00	-.01	.21	-.04	1.4	Dynamic
9	.20	234	234.3	200	.00	.00	.11	.03	.9	Static
8	-.03	96	96.4	70	-.01	.02	.19	.08	.8	Dynamic
2	.43	275	276.9	280	-.01	.02	.09	.17	.9	Static
4	-.13	452	454.2	360	-.01	.02	.08	.19	1.0	Static
11	-.42	280	281.9	200	-.01	.03	.12	.22	1.1	Static
10	-.18	173	174.8	120	-.02	.04	.15	.28	.9	Static
3	.10	526	529.2	360	-.01	.03	.09	.29	1.1	Static
7	-.51	111	112.9	80	-.02	.06	.18	.34	1.0	Static
1	.22	107	110.0	70	-.04	.14	.21	.65	.8	Dynamic
5	-.36	327	335.7	240	-.04	.09	.10	.91	1.1	Static
6	.67	106	125.1	84	-.23	.59	.17	3.50	.4	Dynamic
Mean (count:22)		199.9	199.9	147.3	.00	-.02	.17	-.03	1.0	
SD		124.3	124.3	96.5	.06	.18	.07	1.02	.2	

Note: Fixed (all = 0) chi-square: 23.0 d.f.: 22 significance (probability): .40.

assigned by each rater in each test type (column 11), the total expected score that each rater should have assigned (column 4), the observed count (column 5), and the average value (column 6) between the observed (column 3) and expected (column 4) scores divided by the observed count. The observed score for rater 5 was 119 and the expected score was 109.8. For instance, for rater 5, the obs-exp average (column 6) was .13 (119–109.8 = 9.2/70). Columns 7, 8, 9 and 10 show the bias in logits, SE, t value and the infit mean square respectively. The bias size ranged from -.47 to .59 and the error ranged from .08 to -.33 with a mean of .17, which might be considerable. McNamara (1996) pointed out that the t value should not go beyond the range of +2 to −2. The t value varied from −1.94 to 3.5. In this case, rater 6 in the dynamic test was said to be misfitting with a t value of 3.50. This meant that rater severity varied across test types. In the infit mean square values of raters 6 and 9 in the dynamic test were at the two extremes of the range with .4 and 1.4 respectively. In the observed and expected scores, raters 6, 1 and 8 in static test and raters 5, 11, 10, 3, 7, 4, 2 and 9 in the dynamic test were more lenient than expected. However, raters 8, 1 and 6 in the dynamic part and raters 9, 2, 4, 11, 10, 3, 7 and 5 in the static test were harsher than expected. The fixed chi-square of 23.0 with 22 d.f. was significant at $p < .40$, suggesting that not all the raters were equally severe and that the two test modes were relatively different in terms of difficulty.

Table 5 presents the bias interaction of rater (raters 8, 3, 6, 9 and 1) by test-taker in both test modes. There were 89 instances of biased interactions. Raters 8, 3, 6, and 9 scored candidates 41, 22, 54, 21, 20 and 42 respectively more harshly than expected. The data revealed one case of significant misfit for rater 8 with a t value of 2.17. This meant that rater severity varied across candidates. Also, the infit mean square showed that raters 6, 3, and 1 were at the borderline of the range of misfit; while raters 3 and 9 were identified as misfitting and therefore not consistent in their scoring. The fixed chi-square of 37.4 with 89 d.f. was significant at $p < 1.00$. Therefore, the null hypothesis that all raters were equally severe had to be rejected.

Table 6 shows the bias interaction of test-taker by and test type. There were 120 bias interactions of the total data of 3240. Candidates 23 up to 42 were scored more harshly than expected; thus resulting in a negative value of the Obs-Exp average that ranged from -.38 to -.02. There were cases of less leniently scored candidates, 2, 31, 11 and 27. The SE, column 7, ranged from .24 to .74 with a mean of .38. This suggested that the SE span was large. In observing the t value, there were 7 cases of significant fit all of which stemmed from the dynamic test with candidates 23, 53, 31, 52 and 41 with a significant misfit of 3.22, 2.60, 2.01, 2.30 and 2.17 respectively and other cases of significant overfit, such as candidates 11 and 27 with a value of −2.02 and −2.12 respectively. The bias between candidates and test type indicated that the candidates' ability varied across test type and across test items. In the infit mean square, there were 7 cases of significant misfit having a value beyond the range of 0.4 and 1.60 in the dynamic test for candidates 23, 53, 52, 51 and 20 and candidates 2 and 31 in the static test. There were also 3 other cases that verged the borderline of the range for candidates 55, 21 and 42. The chi-square was 122.7 with 120 d.f. was significant at .42. Thus, the candidates' ability in both test modes differed significantly.

Findings from the interview

This section targeted the test-takers' and raters' perceptions of both modes of assessment. It attempted to answer the following question:

c. What are the mediators' and test-takers' perceptions of both modes of assessment?

Table 5 Bias analysis of rater and test-taker

Rater	Severity	Obsvd Score	Exp. Score	Obsvd Count	Obs-Exp Average	Bias Size	Error.	t	Infit MnSq	Candidate
8	-.03	18	22.5	14	-.32	.90	.41	2.17	.5	41
3	,10	24	26.0	14	-.15	.82	.55	1.49	.9	22
6	.67	18	19.3	14	-.09	.23	.41	.56	.4	54
3	.10	44	46.2	40	-.06	.13	.24	.54	.4	21
9	.20	18	18.7	14	-.05	.12	.41	.29	1.9	20
8	-.03	17	17.2	14	-.02	.04	.41	.10	.4	42
3	.10	72	72.0	54	.00	.00	.22	-.01	1.7	29
1	.22	88	88.0	54	.00	.00	.26	.00	1.6	2
Mean (count:89)		49.4	49.4	36.4	.00	-.02	.32	-.01	1.0	
SD		23.9	23.8	16.6	.09	.30	.13	.65	.3	

Fixed (all = 0) chi-square: 37.4 d.f.: 89 significance (probability): 1.00.

Table 6 Bias analysis of Test-taker and test type

Candidate	Ability	Obsvd Score	Exp. Score	Obs-Exp Average	Bias Size	Model S. E.	t	Infit MnSq	Test type
23	2.13	20	25.3	-.38	1.40	.44	3.22	1.8	Dynamic
53	1.65	16	21.7	-.41	1.05	.40	2.60	.3	Dynamic
31	2.40	23	25.9	-.21	1.02	.51	2.01	.9	Dynamic
52	1.36	15	20.3	-.38	.92	.40	2.30	.2	Dynamic
41	1.14	18	22.5	-.32	.90	.41	2.17	.5	Dynamic
55	1.46	17	20.8	-.27	.69	.41	1.69	.4	Dynamic
51	1.23	16	19.6	-.26	.62	.40	1.55	.3	Dynamic
54	1.17	18	19.3	-.09	.23	.41	.56	.4	Dynamic
21	.60	44	46.2	-.06	.13	.24	.54	.4	Static
20	.59	18	18.7	-.05	.12	.41	.29	1.9	Dynamic
42	.11	17	17.2	-.02	.04	.41	.10	.4	Dynamic
2	1.89	65	63.7	.03	-.12	.30	-.39	1.8	Static
31	2.40	73	70.1	.07	-.41	.40	−1.01	1.7	Static
11	.69	23	17.9	.36	−1.03	.51	−2.02	.8	Dynamic
27	.89	26	20.8	.37	−1.57	.74	−2.12	1.0	Dynamic
Mean (count: 120)		36.6	36.6	.00	-.02	.38	.02	1.0	
SD		17.5	17.2	.14	.46	.13	1.01	.3	

Fixed (all = 0) chi-square: 122.7 d.f.: 120 significance (probability): .42.

In their perception of the dynamic test, 6 out of 11 mediators showed that they managed to organize the turn-taking among the test-takers; while 5 out of 11 pointed out that one of the test-takers dominated the conversation. Even though the mediators (9 out of 11) found that working dynamically on the test was useful and appropriate for their test-takers, 8 out of them agreed that they faced difficulties. All of them maintained that the test-takers were familiar with the static version of the test only, and, therefore, suggested that it was preferable to design, administer, and score traditional static tests. Scoring dynamic listening was likely to be subjective and not practical. Six of the 11 mediators replied that exposing students to both modes of assessment would be a better alternative for assessing LC ability, though they remained doubtful about scoring the test-takers' ability in an objective way. Some of them (n = 7) argued that it was difficult to score their performance in the dynamic test, they did not tend to tolerate the test-takers' grammar, pronunciation and coherence problems. Others added that using DA is not fruitful on the grounds that students at university always tended to be passive in many courses. However, two out of the 11 mediators argued that DA could help learners overcome their language problems.

Concerning the static part of the test, the raters (n = 8) justified that the most given variety of question types, whether in teaching or testing listening, was the MC, wh-, yes/no items. One of the questions given in the static part of the test was on following instructions. Four teachers strongly agreed that such a test item was common to work on only in class and not in exams. A mismatch was found between what was done in teaching and testing. Probably, the listening teachers stuck to the questions in the textbook. In this study context, all the mediators agreed that most often

testing listening was designed with three main parts in the exam: True/False statements, wh-questions and a third part dealing with gap filling (Hidri 2010b).

As for the test-takers, 80% of them noted that the division of the test into 3 parts was very helpful, as they gradually felt more motivated. Generally, they claimed that the interaction with their partners made them more relaxed and that they preferred to sit for similar tests as official exams. Some test-takers (n = 18) assured that they liked to interact with their colleagues in the exam to have good marks. Although 90% of the respondents reported that the static test was more difficult than the dynamic one, nearly 67% of them indicated that they preferred to be tested in a static way basically for practical reasons. That is, they strongly agreed that their partners dominated the conversation. Some test-takers felt nervous in the dynamic test as they were not familiar with some mediators who, in some instances, did not manage to engage them to interact with their partners. However, 27% pointed out that the mediators dominated the conversation, and, therefore, influenced their answers. Others ($n = 15$) noted that the mediators were not helpful, since they did not allocate them enough time to finish their tasks.

All the respondents reported that they were never tested in a dynamic way as 100% of them agreed that the only variety that was given to them was the static classical version. Many test-takers (n = 41) suggested to sit for both test modes to have a comprehensive view about their listening ability, while few test-takers (n = 8) preferred to work on the test individually. All the test-takers reported that in class they were familiar with the questions of the pre-, while-, and post- testing phases, with the exception of using pictures to summarize a story. Seventy six percent of the test-takers agreed that their answers reflected their language ability in English while 20% reported the opposite, because their partners dominated the conversation. In addition, most of the test takers (56%) agreed that they were both familiar with the question types of the static test. Some test-takers (n = 21) assured that they felt nervous in the static test, mainly because exams for them generally entailed stress and anxiety. As for the kind of problems the test-takers had in both modes of assessment, the raters emphasized that the test-takers generally faced some difficulties which were related to making the appropriate inference, grammar, comprehension, and appropriateness and relevance of the answer.

Conclusions

The purpose of the study addressed a need to examine and improve current assessments of listening of Tunisian university EFL test-takers. The study addressed the necessity to use DA in this context and at the same time it explored the classical mode of assessing LC. To this end, different methods of data collection (FACETS analyses and interview data) were utilized. DA generally proved to be a more effective mode of learning. Results of the study confirmed the findings of other studies (e.g., Gibbons 2003; Poehner & Lantolf 2005) when they concluded that learners perform better in joint activities. This finding is echoed in the studies of Gibbons (2003) and Lidz (2002) who maintained that when learners are engaged in a joint activity, they can be very helpful and insightful not only to overcome the difficulty of the test items but also to reach the stage where they can construct

meaning in an autonomous way. This study on assessing listening dynamically yielded the following:

- There was an impact of the mediators' use of support and guidance on the students' processing.
- In some instances, the mediators' lack of support in the post-testing phase resulted in poor performance on the part of the test-takers.

First, this impact was shaped by the teaching experience, views of language and language learning and involvement in and perception of DA. Second, apparently, some mediators (n = 4) tended to score the test-takers' pronunciation rather than appropriateness of the answer. For instance, some mediators could not tolerate grammar and pronunciation problems and, therefore, behaved accordingly (e.g., raters 2 and 6) even though in the benchmarking sessions they were advised not to penalize students for such language problems. Finally, the mediators' less degree of involvement in the post-testing phase did not help the learners to process the task.

DA practitioners have called for the necessity of test-takers benefitting from each other. Yet, the interview feedback showed that generally the test-takers did not benefit much and they did not even benefit from the mediators' support. This was in part due to the fact that mediators were not successful in organizing the turn-taking. It might be important at this level to consider the teachers' roles in class in this particular context and to investigate perceptions of DA in helping the learners build up an independent learning behavior. Generally, dynamic testing can be beneficial in making good progress in learning. However, from the instances of interaction observed, there were occasions where learning did not take place, especially when the interaction amounted to a particular type of dominance, like expert/novice or high versus low proficiency level students. This led to different scores. Therefore, results of the FACETS analyses indicated the following:

- Generally, the test-takers' ability estimates varied significantly in both test modes, with more able students in the dynamic than in the static test. This high performance might be due to the accessibility of the test items, the lenient scoring behavior and the joint interactions.
- The raters' behavior changed depending on the nature of the test in that the scoring resulted in significantly higher scores in the dynamic than in the static test. In fact, this reflected the raters' views of language and language learning.
- The raters behaved more harshly in the static test but were consistently lenient in the dynamic test. This was echoed in the negotiation of meaning.

Although some raters, those who had a longer experience in teaching, had a higher level of inter-rater agreement, they, nevertheless, did not have intra-rater agreement. It is vital for the raters to undergo an intensive and continuous training in order to reduce the measurement inconsistency. Another major discussion point worth mentioning is the use of qualitative data through interviews. This use was very beneficial in probing into the main attitudes towards both types of assessment. These instruments significantly helped probe into the realities of classroom teaching, learning and assessment.

DA may loom beneficial for learners who are mediated to activate their cognitive and metacognitive strategies to notice things. In classical standardized testing, however, such mediation is not offered. DA may be at stake when validity and reliability are concerned. These two notions have been largely addressed in psychometric standardized testing. However, DA researchers have not managed to find reasonable arguments for validity and reliability, except for Lantolf (2009) and Poehner (2011). In this study, DA was not reliable in that when the same measurement procedures were repeated they did not produce the same results, given the fact that the mediation context changed from one learner to another and from one mediator to another. Lantolf (2009), p. 365 argues that "DA makes a strong claim with regard to predictive validity." DA focuses on changing the learner to better levels of linguistic attainment. Since the use of effective dynamic instructions leads the test-takers to perform better in the future, proponents of DA (e.g., Lantolf & Poehner 2009) point out that this future success does in fact echo predictive validity. Contrary to such studies, the test-takers in this study performed well with mediation, but once they were left alone or once the mediators reduced help, they were indecisive and unable to continue processing the test items. Engaging all the raters in training sessions might minimize rater inconsistency and possibly reach objective scoring. Yet, if some of the raters had a more or less similar experience in teaching and were involved in regular training sessions, the results of the study might be different.

Implications

This study addressed a need to examine and improve current assessments of LC. It had theoretical, pedagogical and methodological implications which could be addressed for future research. First, in the theoretical implications, results of DA brought to light the fact that there should be an interface between language learning and language testing. This interface has been addressed in research (e.g., Alderson 2005; Bachman 1989; Bachman & Cohen 1998; Douglas & Selinker 1985). This link integrates instruction and assessment in class to help the learners meet their needs and reach the stage where they can perform independently. DA is not an alternative to classroom assessment, nor can it replace other types of assessment. Rather, it is integrated with classroom instructions to help test-takers overcome their testing difficulties by, for instance, developing their cognitive and metacognitive processes. The findings of DA interactions can be considered additional contributions to the link between assessment and learning. Like other DA studies (e.g., Ableeva 2008; Ableeva & Lantolf 2011; Gibbons 2003), this study showed that with supportive interactions, for instance, in the pre- and while-testing phases, effective learning can take place and that targeting the activation of the learners' cognitive and metacognitive strategies to overcome the testing difficulties.

Second, the pedagogical implications addressed the different steps through which teaching and testing can be improved. In this regard, assessing the learners in a progress dynamic test can help locate the areas of weaknesses in the language program or in the learners' cognitive and metacognitive strategies. Additionally, this assessment can target the measurement of static listening as a final achievement test. In addition, grabbing the test-takers' attention to notice things and praising them to overcome their difficulties are in fact at the heart of any learning process. Research on DA and learning

in general highlights this endeavor. Despite the threats to validity and reliability of the test, assessing learners in a dynamic way in the Tunisian context may be practical and useful given the tremendous language problems these learners have. In terms of authenticity, DA echoes the authentic tasks and activities that the learners are supposed to meet in everyday life, not like psychometric standardized tests. In short, implementing DA has the goal of changing the learners' behavior in their perception of the different courses undertaken at the university level in Tunisia.

Third, the methodological implications called for the importance of using qualitative (interaction in the dynamic test and interview) and quantitative instruments (test scores). Like other studies (Buck 1994), the use of qualitative and quantitative methods played a crucial role in assessment. The feedback teachers suggested about the nature of problems has immediate implications for teaching as well as for testing. In the light of this feedback, the teachers can address and remedy these shortcomings in teaching, and, therefore, in testing.

Limitations and Directions for Future Research

Scoring the joint performance of the dynamic test in an objective way posed many challenges for the mediators who tended to be more lenient in the dynamic than static test. The mediators' kind of interaction with test-takers varied considerably from one teacher to another and, therefore, resulted in different scores. The scoring led to inter-rater and intra-rater agreement in terms of leniency in the dynamic test and inter-rater and intra-rater agreement in terms of severity in the static test. This in fact had serious threats to the validity and reliability of the test. In addition, unlike other studies on DA, (e.g., Gibbons 2003; Lidz 2002), which were carried out through a four- to five-week period of time, this study was carried out in a shorter period of time. While dynamic learning stresses the idea of joint interactions between learners, it fails to account for the pauses of silence where learners produced no output. That is, it may appear hard to find explanations for the silence instances of the learners and to claim whether they were signs of language processing or language problems. Other listening passages, other well-trained raters who were familiar with DA, other candidates with a different ability, another rating context, and other educational and research contexts might have probably led to different results, yet not very divergent from the ones outlined in this study.

There are possible orientations which can be considered for future research. First, there is a need to investigate why raters tended to score dynamic performance more leniently. This could be addressed by investigating the rating experience, views of language and language learning and assessment and their impacts on test scores. Much more qualitative research can be carried out on the paired interactions of the test-takers in DA throughout a longer period of time. Using think-aloud protocol to probe into these silent instances might possibly yield more insights into the nature of acquiring and processing listening. A further investigation into the nature of joint interaction, whether in teaching or testing, and why mediators tend to be more lenient in DA, should be highlighted and investigated. Teachers should address the dynamic nature of tasks, whether in teaching or testing LC, and they should be encouraged to teach listening. Adhering to testing LC dynamically should be highlighted in the Tunisian context

given the poor test scores of the candidates in this skill and in the other language skills. This is a research gap which needs to be addressed to tackle the matching between teaching and testing through using DA. Overall, standardized tests are limited in uncovering the cognitive strategies of learners. At the same time, DA may put the validity and reliability of the test at stake. Hence, assessing the LC ability using both assessment modes might be important in reaching fair inferences on this ability.

Endnotes

[a] The Baccalaureate exam is a compulsory national exam administered to all students who finish their secondary education. It is the equivalent of the A-level. Students can specialize in one of these disciplines three years before they sit for the baccalaureate exam: Arts, mathematics, experimental sciences, economics, technical sciences or information technology.

[b] Scores below 9.99 for these disciplines are the following: Arts, 77.67 (with 40.47 who got below the score of 4), mathematics (55.64), experimental sciences, 67.31, economics 94.81 (with 52.45 who got below 4), technical sciences, 79.58 and information technology 88.97.

[c] The LMD is a newly implemented educational system in the Tunisian universities that dates back to 2006/2007. It consists of reducing the number of study years from 4 to 3. This was done on the assumption that it would minimize cost effects and align the Tunisian educational system with the European ones, since Tunisia has been receiving funds from Europe. Each university is responsible for designing and implementing its own course degree that has to be approved by the Ministry of Higher Education. Since 2006/2007, students, teachers, parents and some policy makers have been complaining about the low level of the language ability of graduates of English. Still, no po.

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

Authors' contributions
All authors read and approved the final manuscript.

Acknowledgements
I would like to thank the three anonymous reviewers for their invaluable feedback. I, however, remain fully responsible for the contents of this article.

Received: 24 February 2014 Accepted: 17 April 2014
Published: 2 May 2014

References

Ableeva, R. (2008). The effects of dynamic assessment on L2 listening comprehension. In P Lantolf & ME Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 57–86). London: Equinox.

Ableeva, R, & Lantolf, JP. (2011). Mediated dialogue and the microgenesis of second language listening comprehension. *Assessment in Education, 18*, 133–149.

Alderson, JC. (2005). *Diagnosing Foreign Language Proficiency: The interfaces between learning and assessment*. London: Continuum.

Bachman, LF. (1989). Language testing-SLA interfaces. *Annual Review of Applied Linguistics, 9*, 193–209.

Bachman, LF, & Cohen, AD. (1998). Language testing-SLA interfaces: An update. In LF Bachman & AD Cohen (Eds.), *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Berne, LE. (1995). How do varying pre-listening activities affect second language listening comprehension? *Hispania, 78*, 316–329.

Bond, TG, & Fox, CM. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences. mahwah*. NJ: Lawrence Erlbaum.

Brindley, G, & Slayter, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*(4), 369–394.

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11*(2), 145–170.

Buck, G, & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157.

Canale, M, & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47.

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

Douglas, D, & Selinker, L. (1985). Principles for language tests within the 'discourse domains' theory of interlanguage: Research, test construction and interpretation. *Language Testing, 2*, 205–226.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23–51.

Gass, S. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Lawrence Erlbaum.

Gibbons, P. (2003). Mediating Language Learning: Teacher Interactions with ESL Students in a Content-Based Classroom. *TESOL Quarterly, 37*(2), 247–273.

Ginther, A. (2002). Context and content visuals and performance on listening comprehension stimuli. *Language Testing, 19*(2), 133–167.

Helal, F. (1997). *Error treatment in Tunisian EFL classes: An application of the communicative competence model*. Tunisia: Unpublished DEA thesis. University of Manouba.

Hidri, S. (2010a). Comparison of students' performance in dynamic vs. static listening comprehension tests among EFL learners. In *paper presented as work in progress in the 32nd Language Testing Research Colloquium, April 14–16, 2010 at the University of Cambridge, Crossing the threshold levels, domains and frameworks in language assessment*.

Hidri, S. (2010b). Writing listening comprehension test items and tasks for learners of English at the tertiary level: Biasing for the Test. In *paper presented at the International Conference on English Language Teaching and Testing: Developments and Challenges, 22–23, April 2010, at the Higher Institute for Applied Studies in the Humanities, Zaghouan, Tunisia*.

Hidri, S. (2013a). Assessing Static vs. Dynamic Listening: Validation of the Test Specifications. In *Published PhD Dissertation*. LAP LAMBERT Academic Publishing.

Hidri, S. (2013b). The effectiveness of assessment of learning and assessment for learning in eliciting valid inferences on the test-takers' listening comprehension ability. In *article published in the Proceedings of the Nile TESOL Conference: Revolutionizing TESOL: Techniques and Strategies* (pp. 1–25). Egypt: The American University in Cairo. https://docs.google.com/file/d/0B6bmHwcjFuVYX2FJZTFndGF0QzA/edit.

Hidri, S. (2014). Comparison of the students' performance in dynamic vs. static listening comprehension tests among EFL learners. In *article published in the Proceedings of the 19th TESOL Arabia Conference, From KG to College to Career* (pp. 51–59).

Jensen, C, & Hansen, C. (1995). The effect of prior knowledge on EAP Listening-test performance. *Language Testing, 12*, 99–119.

Karpov, YV, & Haywood, HC. (1998). Two ways to elaborate Vygotsky's concept of mediation: Implications for instruction. *American Psychologist, 53*(1), 27–36.

Kondo-Brown, K. (2002). A *FACETS* analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3–31.

Kozulin, A, & Garb, E. (2002). Dynamic assessment of EFL text comprehension. *School Psychology International, 23*(1), 112–127.

Lantolf, P. (2009). Dynamic assessment: The dialectic integration of instruction and assessment. *Language Teaching, 42*(3), 355–368.

Lantolf, JP, & Poehner, ME. (2004). Dynamic assessment of L2 development: Bringing the pat into the future. *JAL, 1*(1), 49–72.

Lantolf, JP, & Poehner, ME. (2006). *Dynamic assessment in the foreign language classroom: A teacher's guide*. Pennsylvania: CALPER University Park.

Lantolf, JP, & Poehner, ME. (2009). The artificial development of second language ability: A sociocultural approach. In WC Ritchie & TK Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 138–159). Bingley, UK: Emerald Press.

Lantolf, JL, & Poehner, ME. (2010). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research, 15*(1), 11–35.

Leung, C. (2007). Dynamic assessment: Assessment for and *as* teaching. *Language Assessment Quarterly, 4*(3), 257–278.

Lidz, CS. (2001). Multicultural issues and dynamic assessment. In LA Suzuki, JG Ponterotto, & PJ Meller (Eds.), *Handbook of multicultural assessment: clinical, psychological, and educational applications* (2nd ed., pp. 523–539). San Francisco: Jossey- Bass.

Lidz, CS. (2002). Mediated learning experiences (MLE) as a basis for an alternative approach to assessment. *School Psychology International, 23*(1), 68–84.

Lidz, CS, & Gindis, B. (2003). Dynamic assessment of the evolving cognitive functions in children. In A Kozulin, VS Ageev, S Miller, & B Gindis (Eds.), *Vygotsky's educational theory in cultural context* (pp. 99–116). New York: Cambridge University Press.

Lumely, T, & McNamara, TF. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71.

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

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

McNamara, T, & Roever, K. (2006). *Language testing: The social dimension*. Oxford: Blackwell.

Ohta, A. (2000). Rethinking interaction in SLA: Developmentally appropriate assistance in the zone of proximal development and the acquisition of L2 grammar. In JP Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 51–78). Oxford: Oxford University Press.

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

Poehner, ME, & Lantolf, J. (2005). Dynamic assessment in the language classroom. *Language Teaching Research, 9*, 233–265.

Poehner, ME, & van Compernolle, RA. (2011). Frames of interaction in Dynamic Assessment: Developmental diagnoses of second language learning. *Assessment in Education: Principles, Policy and Practice, 18*(2), 183–198.
Rea-Dickins, P. (2006). Currents and eddies in the discourse of assessment: A learning-focused interpretation. *International Journal of Applied Linguistics, 16*, 164–189.
Rubin, J. (1994). A review of second language listening comprehension research. *Modern Language Journal, 78*, 199–221.
Sternberg, RJ, & Grigorenko, EL. (2002). *Dynamic testing. The nature and measurement of learning potential*. Cambridge: Cambridge University Press.
Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through dynamic dialogue. In JP Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97–114). Oxford: Oxford University Press.
Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275–282.
Tzuriel, D. (2011). *Revealing the effects of cognitive education programmes through Dynamic Assessment, Assessment in Education: Principles, Policy & Practice* (Vol. 18, pp. 113–131).
Vandergrift, L, Goh, CCM, Mareschal, CJ, & Tafaghodtari, MH. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language Learning, 56*(3), 431–462.
Vygotsky, L. (1981). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University Press.
Vygotsky, L. (1986). *Thought and language*. Cambridge, MA: MIT Press.
Yi'an, W. (1998). What do tests of listening comprehension test? A retrospective study of EFL test-takers performing a multiple choice task. *Language Testing, 15*(1), 21–44.

doi:10.1186/2229-0443-4-4
Cite this article as: Hidri: Developing and evaluating a dynamic assessment of listening comprehension in an EFL context. *Language Testing in Asia* 2014 **4**:4.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com