Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 11, No. 1, March 2021
ISSN 2476-5880

Assessment Alternatives in Developing L2 Listening Ability: Assessment FOR, OF, AS Learning or Integration? Assessment x̄ Approach

Elham Ghorbanpour[1], Gholam-Reza Abbasian[2]*, Ahmad Mohseni[3]

Received: 9 November 2020 Accepted: 10 February 2021

Abstract

Uni-furcation of assessment and instruction has recently been realized in the form of purposeful assessment scenarios; Assessment x̄ Scenarios (analogous to Noam Chomsky's ¯ Theory!). ¯ here refers to any of the triple assessment scenarios including Assessment for Learning (AFL), Assessment as Learning (AAL), and Assessment of Learning (AOL), plus pairing each with another or integrating all three (i.e., Integrated Assessment Scenario). Comparative investigation of the effect of each scenario as to developing language skills particularly listening skill seems to be an intact area. In a bid to fill this gap, 100 conveniently sampled Iranian female EFL learners of 13-19 years old were randomly divided into three experimental and one control group. Prior to the treatment, their listening ability was measured through a pre-test. Then, each experimental group; AFL, AAL, and Integrated assessment, experienced the listening instruction based on the principles of each specific scenario, while the control group was treated based on AOL principles. Their listening ability was then measured in the light of an identical listening post-test to the pre-test. ANOVA, used to check the comparative performances of all groups, showed that AFL and AAL groups significantly outperformed over the AOL group, but the integrated assessment group significantly outperformed the other experimental groups. While the findings yield support to the bifurcation approach, they generate more prospective areas for further research.

Keywords: Assessment; Listening Ability; Purposeful Assessment Scenarios

1. Introduction

Assessment and evaluation are essential components of teaching and learning in English language arts. Without an effective evaluation program, it is impossible to know whether

[1] PhD Candidate, Faculty of Persian Literature and Foreign Languages, Islamic Azad University- South Tehran Branch, (Kish Int'l Campus), Tehran, Iran; elhamghorbanpour4@gmail.com
[2] *,Assistant Professor, English Language Department, Faculty of Basic Sciences, Imam Ali University, Tehran, Iran, (Corresponding author): gabbasian@gmail.com
[3]Associate Professor., English Language Department, Faculty of Persian *Literature* **Foreign** Languages, Islamic Azad University- South Tehran Branch, Tehran, Iran; Amohseny1328@gmail.com

students have learned, whether teaching has been effective, or how best to address student learning needs. However, language assessment as an important part of the English language teaching puzzle has not fully fallen into place as many teachers are still underprepared for bridging language assessment research-practice gaps in classrooms (Babaii & Asadnia, 2019). Though scholars like Marlone (2013) dealt with the gap and it is widely agreed that classroom teachess eeed to ass–ss steeett s' ggggress, many teaceess add test ssess aave a limited understanding of assessment fundamentals. Teachers usually consider assessment as an evaluation tool rather than as a learning instrument or quality learning to promote the steeett s' eeeeiicial leaiii gg MMii ssette, 2.11.. . eey aee eeeaally meee ccccceeeed with measuring achievement in both the summative and formative senses of the term. In these kinds of assessment, there is a great emphasis on comparing the learners, and feedbacks come in the form of grades or marks, with little advice or direction for improvement. These types of assessments reveal which learners are doing well and which ones are performing poorly. Generally, "they don't give much indication of mastery of particular ideas or concepts because the test content is generally too limited and the scoring is too simplistic to represent the broad range of skills and knowledge that have been covered" (Earl, 2013, p. 29).

It seems that these types of assessments are not very useful for communicating cmmll ex aata atttt a stddett ss aii lities (Fll ceer & Daviss aaaa2..... . he same coccenn was raised by Popham (2004), who termed the lack of appropriate training in the assessment as "professional suicide" (p. 82).

Generally, assessment of learning (AOL) and grading has a long history in education and it is the predominant type of assessment in the Iranian academic context. Almost many classroom assessments in a traditional environment are summative or AOL focused on measuring learning and categorizing students and reporting these judgments to others. Thus, numerous educational researchers and theorists have discussed the traditional grading practices for quite some time, and the area is also well researched in the Iranian setting. Assessment for learning (AFL) offers alternative perspectives to traditional assessment; however, a few teachers use AFL for diagnostic purposes and give students feedback to improve their marks and their learning in the Iranian EFL context. Moreover, Lee (2016) identifies dominant traditional assessment paradigms, large class sizes, and students' inadequate linguistic proficiency as major challenges to the use of assessment as learning (AAL). These tensions have turned AAL into an empirically unexplored area with narrow application awaiting comprehensive attention and clarification (Lee, 2016). Moreover, research and explanations about how AAL can be implemented in the L2 listening classroom are scant in general and almost nonexistent in the Iranian EFL investigations.

Although the literature indicates that these three approaches (i.e., AOL, AAL, AFL) all contribute to student learning, most of the time the emphasis is on a particular type of assessment not the integration of these three approaches. It also seems that the integration of assessmett aaaaaaaees aas yieleed ii fferett imaacts nn tee steeett s' academic acii evemett in different contexts. For instance, some scholars (e.g., Brookhart, 2001; Carless, 2011; Earl, 2003; Lam, 2013) argued that the integration of assessment approaches could promote the students' achievement. Likewise, in the Iranian context, Sadeghi and Rahmati (2017) also proved that the integrated assessment group (AOL, AAL, AFL) outperformed the non-

integrated assessment participants. However, other researchers, (e.g., Harlen, 2006; Lam & Lee, 2010; Lee & Coniam, 2013; Taras, 2005; Wei, 2015) agree that the integration of different approaches was not much effective.

Contrary to the fact that most of the research on L2 listening has focused on improving listening comprehension, methodologies to teach listening, the appropriateness of listening materials, and similar matters, few have focused their attention on the assessment of listening comprehension in general and integrating different assessment methods to evaluate this skill in particular. There is almost no research that examines the synergistic effect of assessment instruments on listening comprehension in Iran in addition to the above-mentioned challenges.

To sum up, despite the vital importance given to the assessment of listening skills, there are still some gaps, especially in the Iranian context. It seems that there is a marked absence of research on the integration of different assessment approaches in Iran. Consequently, the present investigation tried to address the gap in the previous research studies by uncovering the effects of integrating various assessment approaches, namely AOL, AALa add AFLt tt t tt Iaaii aE EFL leaeeesse eevelmmmatt ll l lsteii gg a55-ty.

2. Review of Literature

2.1. Language Assessment

Assessments are generally considered as a tool for informing the teachers and the learners regarding their practice in the classroom, and what they are required to do in facing challenging issues for academic success (Stiggins, 2002). One of the important roles of EFL teachess in the laggaage classmmmmis to assess steeett s' eemmmmmmce (Watmaii , Asadollahfam & Behin, 2020). When it comes to the assessment of language skills, instructors are concerned with selecting the most practical and applicable approach to evalaate tee leanness' ggggeess add ccccett aate nn their ooitt s ff steeggth add weaeee ss. According to Drummond (2003), in the assessment process, teachers collect and interpret evieecce ff steeett s' leaiii gg add alll y taat immmmmmn to maee decisisss . The psssss ss ff such decisions are of significant importance in determining the function and purpose of assessment and may influence the choice of assessment methods. James (2013) supported this by emphasizing that fitness for a purpose is a comprehensive principle that should guide all assessment practices.

As an innovative breakthrough in education and in line with the unification approach, AFL, AAL, and AOL emerged when constructivism (Piaget, 1960) attempted to pinpoint the role of assessment in teaching language skills and whether the three above-mentioned assessment approaches could facilitate tee leaeeessd develmmmatt ff tee laggaage sii lls.

2.1.1. Assessment of Learning (AOL)

Assessment related to behaviorist perspectives attempts to test whether the students have met the set requirements (James, 2006). This type of assessment characterizes AOL (Berry, 2008). AOL is utilized to plan learning targets of students and provides evidence of achievement to

the broader community. Sadeghi and Rahmati (2017) highlighted a few of the issues that are currently controversial about AOL. Although there has always been a large support for grading in schools, there is a growing skepticism regarding its accuracy and fairness. Educational theorists and researchers have criticized traditional grading practices for a long time. Grades are greatly suspect in terms of measurement theory. The reason is that when teachers assign grades, they consider multiple factors other than academic achievement; they weigh assessments differently, and they make a misinterpretation that a single score on assessments can reflect performances on a wide range of abilities and skills (Marzano, 2000). Moreover, according to Huang (2012), being decontextualized, one-shot, indirect, and product-oriented, with no clear mechanisms for delivering constructive and helpful feedback are the criticisms against these types of assessments.

2.1.2. Assessment for Learning (AFL)

Throughout the 1980s, formative assessment (FA) and Assessment for Learning (AFL) (Earl, 2013; Lam, 2013) offered an alternative perspective to the psychometrics period and the traditional assessment in schools. This type of assessment, which emphasizes the assessment potential to support learning, has gained considerable attention in educational settings over tee last eecaees EEarl &Timeerldy, 2014.. Fmmative assessmett is "all tsss e activ4ies aee undertaken by teachers, and/or by their students, which provide information to be used as eeeaaack to miii yytee teacii gg add leannigg activities in wii ch -eey aee eggage"" BBack & Wiliam, 1998, p. 7).

The philosophical principle of AFL considers learners as agents in their learning. In much literature, this approach is typically referred to as "formative assessment". This term is related because it focuses on a similar philosophical principle; that is, using evaluation "as a learning tool" rather than for certification (Glazer, 2014, p. 277). AFL practices faced several challenges at the school level. The greatest challenge for the implementation of formative assessment is the accountability issue (Broadfoot, Oldfield, Sutherland, & Timmis, 2014). Pham and Renshaw (2015) concluded that in Asian settings, the implementation of AFL was hindered by the potential discrepancy between AFL principles and the local learning culture. For instance, peer assessment often needs learners to discuss with their peers; however, the Asian countries' culture might make learners reluctant to challenge their friends' opinions as well as to evaluate their work. Therefore, although the culture of AFL was encouraged in many universities, Medland (2014) states that there is much criticism of assessment practice in this setting because of the persistence towards a testing culture. Such findings indicate that AFL encounters many tensions in all higher education environments.

Bayat, Jamshidipour, and Hashemi (2017) examined the influence of using formative assessment on EFL students' listening efficacy and anxiety. Instructors using formative assessment reported that the learners retained more information, understood concepts more quickly, and were more interested in what they were learning. The teacher investigated to find out weeteer the sse ff mmative assessmett aad ayy imaact nn tee leanness' listeii gg enhancement. Indeed, there were substantial differences in the achievement level of the experimental group as regards listening efficacy and comprehension compared to the summative control group.

In an experimental investigation, Zarei and Yasami (2016) examined the impact of formative assessment and remedial instruction on Iranian EFL students' listening comprehension. Data analysis indicated that formative assessment and remedial instruction had a considerable impact on the listening comprehension of EFL students.

2.1.3. Assessment as Learning (AAL)

Tee cmnbiaatinn ff leaiii gg add teacii gg can valee tee leaeeess' eeesecce in the assessmett of the learning progress, leading to the emergence of AAL (Dann, 2014; Lam, 2015). In AAL, students are engaged in the assessment process, which can enable them to become responsible for their learning behaviors, leading to their self-reflection and self-monitoring (Archer, 2010). In the AAL, learners are their assessors, which results in a learning-oriented procedure through which the learners are expected to manipulate the learning environment (Lam, 2015). According to (Lee, 2016) some major challenges to AAL practice used in the classrooms are large classes, dominant traditional testing paradigms, and students' low linguistic proficiency. Lee (2017) suggested that teachers can adopt the four main strategies to overcome these challenges.

Several investigations in the second language and other fields have explored the implementation of AAL-oriented instruction on the learners' achievement. For example, Xiao and Yang (2019) investigated how formative assessment could support secondary students' self-regulated learning in English language learning. The findings indicated that under the guidance of their teachers, the students proactively engaged in formative assessment and appeared to be emerging as self-regulated learners. In a similar vein, Li (2018) examined the washback and validity of self-assessment, as a specific form of AAL, in interpreter and taass latrr ecccati... The iiddiggs swwwed: tee steeett s' self-assessment correlated positively with their instructor assessment; the assessment accuracy of the students enhanced over time with regular repetition, and self-assessment promoted positive learning attitudes among students.

AAL, AFL, and AOL, as the main approaches towards assessment-oriented instruction (Earl, 2013), are concerned with the frequent themes of time, means, goal, and key factors of assessment as well as who holds the role of assessor (Lee, 2016). Moreover, as argued by Carless (2011), integration of AAL and AFL– as formative assessment - in comparison with AOL – as a summative assessment - can be more useful. Lee (2016, p. 271) also eelieves taat "AOL add AFL/AAL can co-eii st". Howeve,, teeee aave eeen aggmmett s regarding the integration of these approaches. For instance, some scholars (Harlen, 2012; Lam & Lee, 2010; Taras, 2005) agree that when these approaches are integrated, the focus is greatly on summative assessment rather than assessing the learning process by involving the learners. Taking these issues into account, the current investigation attempted to uncover the effect of using different purposeful assessment scenarios of AFL- AAL- and AOL-oriented isstcccctisss in Iaaii an EFL leaeeess' eeeelmmmemt ff listeii gg sii lls. Hecce, tee lll lwwigg research questions were addressed:

Is there any significant difference between the effect of AFL and AOL on L2 listening achievement?

Is there any significant difference between the effect of AAL and AOL on L2 listening achievement?

Is there any significant difference between the effect of (AFL+AAL) and AOL on L2 listening achievement?

3. Method

3.1.Participants and Setting

The participants of the study were 100 pre-intermediate female EFL language learners from a private language institute in Rasht, Iran. They were studying in the Institute when this study was designed to be done and the researchers provided sufficient information about the purpose of the study and what they were to go through in the process of treatment sessions of ssssss slll assessmett sceaariss. Tee aatticiaants' age aagge was eetween 13 to 19 yeaOOHH (M=16, SD=2.02), considering them as young language learners. Concerning the purposes of the study, the researcher organized three experimental groups and one control group. Therefore, four groups were created each of which included 25 participants to meet the purpose of this research. The three experimental groups were exposed to different purposeful assessment scenarios including AFL-oriented instruction, AAL-oriented instruction, and integrated purposeful assessment scenarios (focusing on the combination of adopting the two above-mentioned instructional assessments). Finally, the subjects in the control group underwent AOL-oriented instruction. It is also worth noting that the whole population of the present research was 110 female EFL Pre-intermediate language learners who took the Oxford Placement Test (OPT) to meet the homogeneity assumption of research. Hence, 10 learners were considered as outliers and were removed from the process of selecting the participants, finalizing the total number of them as 100 pre-intermediate young language learners.

3.2.Instrumentation

The instruments applied for data- collection process include:

3.2.1. Oxford Placement Test (OPT)

OPT was used to select a homogeneous sample of the participants in terms of their level of proficiency. The test has four parts including grammar, reading, vocabulary, and writing, and the students were supposed to answer multiple-choice, matching, and cloze tests. Besides, the participants had 60 minutes to complete the test.

The first part consists of three sections. The first section included five multiple-choice questions with three items. The second section contains three cloze tests with 15 test items; each part consists of five questions with four options. The third section consists of 20 multiple questions, and each test item has four options. The second part includes two cloze tests with 10 multiple questions and ten completion questions. The third part is writing, and students are supposed to answer a question in about 150-200 words. The reliability measure of OPT is presented in Table 1.

 ISSN 2476-5880

Table 1
Reliability index of OPT

	aaaaaa asss Alaaa	N of sample
OPT	.81	110 EFL learners

a all e 1 swwws taat tee estimated valee ff tee aaaaaa asss Alaaa rrr the OPT was .81, which can be assumed to be higher than the minimum possible amount required (i.e., .70) as pointed out by George and Mallery (2003) that the reliability coefficient between .80 and .90 is considered good, thus acknowledging the reliability of OPT as a proficiency test.

3.2.2. Researcher-Made Diagnostic Listening Pre-Test

Beeeee essss igg tee aatticiaants to ssssss slll assessmett sceaaiiss, tee leanress' listeii gg comprehension was checked to test their initial knowledge of this skill. Since the participants were of pre-intermediate level, testing their listening comprehension skill was a critical concern for teachers to select the most appropriate, validated, and reliable test for such proficiency level subjects. Thus, the researcher-made diagnostic listening test which was taken from Touchstone Level 1 (McCarthy, McCarten, & Sandiford, 2004) and Basic Tactics for Listening (3rd edition) (Richards, 2013) was administered before the treatment sessions. The pre-test included 36 questions, including multiple choices, matching, and true-false items.

As to the reliability measure of the listening pre-test, a pilot study was conducted with the participation of 75 similar students (from another private institute with similar characteristics to the participants of the present study) to roughly go for the test score consistency. The reliability coefficient was found to be 0.79 (using the KR-21 formula), which seemed to be an acceptable value in terms of consistency of scores as highlighted in Farhady, Jafarpour, and Birjandi (1994). The reliability of the pre-test is shown in Table 2.

Table 2
Reliability index of researcher-made diagnostic listening pre-test

N	Mean	SD	Variance	Reliability
75	21.55	5.99	28.16	0.79

*. The mean difference is significant at the 0.05 level.

Though the content validity of the test was assured through a panel of teachers possessing years of experience in incorporating the sourcebooks in their syllabus, OPT is an instrument that its construct validity has also been checked in terms of apprpriatness for non-natve English language leanrners. Wistner, Hideki and Mariko (n.d.) from Hosei University run a comparative analysis of the Oxford Placement Test and the Michigan English Placement Test. In addition to reprting an acceptable reliability index for OPT, the factor

analysis they run resulted in two factor solution yielding support to the fact that OPT can function as proficiency test for non-native (i.e. Japanese) learners of English.

3.2.3. Researcher-Made Achievement Listening Post-Test

When the treatment sessions of purposeful assessment scenarios were done, there was a need to measeee tee leanress' listeii gg cmnpeeeess inn sii lls add evalaate the effectiveeess ff AFL-oriented instruction, AAL-oriented instruction, and integrated purposeful assessment scenarios. To meet this end, a similar version of the listening pre-test was administered among the participants to investigate their possible improvement in listening comprehension. Similar to the pre-test, the researcher made achievement listening post-test items were eesllll l t t c challegge te ll l eeessc ccc ceeeess s...

Regarding the reliability coefficient of the listening post-test, the same participants, who took part in the pilot study for the pre-test, were considered potential candidates to carry out the post-test to check the consistency of the post-test scores with the application of the KR-21 formula. The reliability was calculated as 0.81 highlighting a logical amount of consistency measure. The reliability of the post-test is shown in Table 3.

Table 3
Reliability index of researcher-made achievement listening post-test

N	Mean	SD	Variance	Reliability
75	23.19	6.11	38.90	0.81

*. The mean difference is significant at the 0.05 level.

It should also be noted that the validity of both pre-and post-test of listening comprehension was checked by three Ph.D. holders of TEFL.

3.3.Procedures

Since the detailed procedures are reported here, this section is divided into: (a) pre-treatment phase focusing on the pilot study, the administration of OPT, and the pre-test; (b) while-treatment phase including the complete explanation of the treatment sessions of the three purposeful assessment scenarios, and the AOL-oriented instruction as the control group; (c) post-treatment phase entailing the information regarding the post-test).

3.3.1. Pre-Treatment Phase

This section has to do with the preparation stage of the research. Initially, a pilot study was done to make sure of the development of sound instruments. Then, the researcher initially held the necessary meetings with the Head of the Institute to make the necessary coordination for data collection procedures. The participants of the study were also provided with sufficient information regarding the purpose of the study. The whole population of 110 language learners took the OPT and 100 subjects, whose scores lied between one SD below the mean, scores were apt to take part in the current research. The selected participants constituted the main subjects for doing all data collection measures during the three-month research. They were also divided into three experimental groups and one control group (i.e.,

AOL-oriented instruction) to meet the purpose of the study. Then, the four groups of learners took the pre-test to check their initial listening skills before exposure to purposeful assessment scenarios, which are explained in the following.

3.3.2. While-Treatment Phase

This section is concerned with the detailed explanation of the treatment session in each experimental group. It is worth mentioning that all four groups of the study were provided with listening activities taken from Touchstone Level 1(McCarthy, McCarten, & Sandiford, 2004) and Basic Tactics for Listening (3rd edition) (Richards, 2013) the two which were covered in their syllabus during the term. Moreover, they all underwent 12 two-hour sessions of instruction and assessmett nn tee leanness' listeii gg eemmmmmce. It is ttt ewttt hy taat all the target purposeful assessment scenarios were consciously applied as a tool in the teaching of listening comprehension. Details of the process are as follows:

3.3.2.1. Purposeful AFL-oriented instruction

This type of instruction aimed to arm the learners with the iterative assessment on their listening assignments. In fact, during the course, the teacher tried to do assessments tgggggggggg tee cssss e to taack –he leareess' listenigg progress. Every two sessions, the learners were required to carry out the listening activities, and do some listening projects as homework assignments and make them ready for the upcoming sessions. Moreover, during the treatment sessions, the teacher tried to interview the participants, in the group, to identify their strengths and weaknesses, and probe their perceptions about the instruction they were exposed to. Assessment for learning was exclusively done by the teacher to look into the leaeeess' eevelppmental progress in their listening and how the existing gaps in the listening instruction could be solved. The teacher attempted to receive feedback in teaching the listening comprehension materials by repeatedly assessing their listening comprehension and scrutinize their listening comprehension progress.

The teacher benefited from both oral and written formats of quizzes (every two sessions quizzes were taken by the participants), questioning, conversations, and learning logs to test their listening development and provide immediate descriptive feedback on their responses. The learners got used to being assessed by the teacher since there was no concern for losing a score at the end of the term or any negative mark. Assessments were done for better teaching add eevelmmmental leaiii ng ff listeii gg cmmeeeeessi... cccc eiii gg teachess' roles, they used AFL-iii ett ed isstccctitt t o eaaa cce steeents' mtt ivatiaa dd cmmmitmett to learning. In this way, they brought changes in the classroom culture by making visible what students believed to be true and used that information to help students move forward in manageable, efficient, and respectful ways. To make AFL more systematic, the teacher benefitted from record-keeping for individual students to provide each student with individualized descriptive feedback that would help further their learning.

3.3.2.2. Purposeful AAL-Oriented Instruction

Tii s tyee ff assessmett was mcch similar to AFL –s tee leanness' listenigg ggggeess was concerned. Integration of teaching and assessing listening comprehension was done in that the

teacher tried to involve the learners in the process of instruction by asse sigg 8heir peess' listening comprehension, which was monitored and supported by the teacher through the provision of feedback during the peer-led assessment. During the treatment sessions, the learners were expected to write diary journals and express their points of view regarding each session. In this way, they were able to have a self-evaluation of their learning behaviors. The teacher attemtt ed to elll rr e add vvvvide itt ssseectinn ff tee leaeeess' metacggii tinn yy allowing them to assess their classmatesc cmmeeeeess inn add caallegge teeir eeee sstaiii gg ff the listening activities they were required to carry out in-class and as homework assignments. In other words, AAL-oriented instruction extends the role of teachers to include designing instruction and assessment that allows all students to think about, and monitor their learning. The learners were also expected to write portfolios concerning the listening activities and any comments they found necessary to be used by the teacher for future classes. This type of assessment, as the name suggests, occurred in alignment with teaching listening and the purpose was to help learners be in charge of their listening success or failure and self-reflect or self-medicate their listening progress during the treatment sessions. No grades were cccc eeeed yy tee teaceer as well. Sicce tee leanress' ieee eeddett leaiii gg is cccc eeeed in AAL, the teacher aimed to teach the learners to have a checklist of their listening activities and focus on their points of strength and weakness, while encouraging them to consult the teacher for having a better image of their listening performance. The main purpose of AAL was to elll eee tee leanress' metacggii tion add eelp teem eecome aware ff it. Ptt it simll y, AAL aimed to directly involve the learners in the assessment procedure to provide efficient and less stressful instruction, which is based on evaluations of their comprehension done by themselves.

3.3.2.3. Purposeful Integrated Assessment oriented- instruction

The third experimental group in this study experienced teacher-centered (AFL) and learner-centered (AAL) assessment scenarios to improve their listening comprehension abilities. Depending on the time of the sessions, the teacher attempted to do the assessment himself and put it on the shoulders of more active learners to benefit from an integrated assessment. However, caution was made by the teacher in order not to bombard the learners with a variety of assessment tools at the same time. A formative assessment procedure was run in this group as tee leaeeess' eevelmmrent ff tee li lteii gg cmmeeeeess inn aii lity was cccss ed eeaalll ess ff their final listening score. The main purpose of assessment in this group was to provide a variety of procedures to help learners be the evaluator and be assessed during the listening comprehension tasks. The former involved writing diaries and portfolios by the learners to be tee teaceesss assistatt in iii gg tee assessmett , wii le tee latter was cccc eeeed with tee teachesss initiatives in playing an imptt att ode in evalaatigg tee leaeeess' listenigg comprehension progress during the course by interviewing them and analyzing the points of strengths and weaknesses.

3.3.2.4. Purposeful AOL-Oriented Instruction (Control Group)

Compared to the three previous assessment scenarios, the learners in this type of assessment were provided with scores and grades at the end of the treatment sessions to check their

overall listening development. During the treatment sessions, an attempt was made to engage learners in the listening activities and provide an interactive listening class. It should be noted that this type of assessment was fulfilled by the teacher-made listening test taken by the participants at the end of the term for summative reports. The process of instruction was conventional and the assessment was mainly done at the end, though process-based and final data were used to make the evaluation. The results of such assessment could be somehow beneficial for the institute, the teacher, and the learners to be aware of their final score regardless of their formative assessment during the term, which was fulfilled by AFL and AAL assessment scenarios.

3.3.3. Post-Treatment Phase

After the treatment session, in the 13th session, the same test of the pre-diagnostic listening test was given to experimental groups and control group as a post-achievement listening test to investigate the effect of different purposeful assessment scenarios on their listening comprehension.

3.4. Data Analysis

To analyze the data, both quantitative and qualitative methodologies were considered to answer the research questions of the study. To analyze the three research questions of the study quantitative methodology including descriptive (mean and standard deviation) and inferential statistics (one- way ANOVA) through SPSS software (version 23) was applied

4. Results and Discussion

4.1. Checking Data Normality Assumptions

To initiate data analysis, normal distribution of data was checked for the four groups to acknowledge the use of parametric or non-parametric tests for data analysis as in Table 4

Table 4
Test of Normal Distribution for the Four Groups

		Shapiro-Wilk		
	Groups	Statistic	df	Sig.
Pre-test	AFL	.89	25	.15
	AAL	.90	25	.18
	Integ.	.79	25	.06
	AOL	.82	25	.12
Post-test	AFL	.90	25	.05
	AAL	.90	25	.07
	Integ	.85	25	.11
	AOL	.82	25	.06

**. The mean difference is significant at the 0.05 level*

As Table 4 shows, p-values of the listening comprehension pre-and post-tests of the AFL group are .153 and .058. The same values for the AAL group include .180 and .071.

Similarly, p-values for the integrated group are .066 and .113. Finally, the same values for the AOL group are .121 and .060. It can be inferred that the p-values for the four groups are more than .05, which indicates that the normality assumption is met. Therefore, as argued by Hatch add Lazaaatnn)1))1))aat "paaametric tests assmme taat tee aata are mmmdly ii stiittt e"" pp 238), the parametric test of one-way ANOVA can be applied for inferentially analyzing the data.

4.2. Homogeneity of Variances

After checking normality assumptions, equality of variances among the groups had to be also checked. Before running the inferential analyses, the homogeneity of the variances was ceeceeu uut tgggggg gmnttt igg Leveeess testw wi ci i s vvvweei i T Tall e ..

Table 5
Levene's Test for Examining the Homogeneity of Variances

Pre-test scores	Levene Statistic	df1	df2	Sig.
AFL pre-test scores	.75	2	97	.21
AAL pre-test scores	.80	2	97	.10
Integ. pre-test scores	.69	2	97	.18
AOL pre-test scores	.73	2	97	.19

**. The mean difference is significant at the 0.05 level.*

Leveeess statistics swwwtaat tee group variances were similar in pre-and post-test scores since p-valees rrr all rrrr rssssss listeii gg sceees weee meee than .... Leveeess statistics supported the hypothesis that the group variances were the same for the pre-and post-test-test scores, justifying the parametric inferential tests applied in the study.

The three research questions of the study are taken into account below:

4.3. Addressing the First Research Question.

The first research question aimed to scrutinize the impact of purposeful AFL-oriented teachigg nn tee leaeeess' listeii gg cmmeeeeess inn sii lls. To oo s,, the aaatt itative aaalysis ff tee leannesse eee-and post-test scores was run. In Table 2, the descriptive statistics of AFL and AOL groups for the post-test are indicated.

Table 6
Descriptive Statistics for the Post-Test of Three Groups

	Groups	N	Mean	Std. Deviation	Std. Error Mean
Post-test	AFL	25	31.00	3.99	.99
	AAL	25	30.50	3.98	.98
	AOL	25	28.00	4.52	1.19

**. The mean difference is significant at the 0.05 level.*

Descriptive statistics for the post-test listening comprehension of AFL and AOL shows a noticeable difference between the two since the AFL group (M= 31.00) performed better than the AOL group (M=28.00).

Table 7 presents the inferential statistics for the listening comprehension scores of AFL and AOL on the post-test.

Table7
One-way ANOVA Statistics for the Post-Test of Three Groups

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	421.11	2	233.76	7.11	.00
Within Groups	1341.36	72	27.81		
Total	1762.47	74			

*. The mean difference is significant at the 0.05 level.

As to Table7, it can be concluded that there exists a significant difference among the listening comprehension post-test of three groups (F2, 72= 7.11, p= .003). Thus, it can be inferred that the three groups were significantly different in their listening comprehension gain, which is an indication of rejecting the null hypothesis. To locate and highlight the difference among the three groups, Tukey HSD Multiple Comparison was run as shown in Table 8.

Table 8
Tukey HSD Multiple Comparison Statistics for the Post-Test of Three Groups

	(J)	Mean			95% Confidence Interval	
(I) Groups	Groups	Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
AFL	AAL	0.50	1.01	.12	.98	9.06
	AOL	3.00	1.43	.00	3.79	9.34
AAL	AFL	-0.50	1.01	.12	-9.06	-.98
	AOL	2.50*	0.99	.00	2.84	7.66
AOL	AFL	-3.00	1.43	.00	-3.79	-9.34
	AAL	-2.50*	0.99	.00	-2.84	-7.66

*. The mean difference is significant at the 0.05 level.

The table shows that a significant difference can be observed between AFL and AOL groups (p= .001) and AAL and AOL groups (p= 002), while no significant difference is seen between AFL and AAL groups (p= .129). In other words, there was a significant difference between the effects of purposeful AFL-oriented instruction and AOL-oriented instruction in developing Iaaii an EFL lea--- ss' listeii gg cmmeeeeessi... S,, we can eeject the lll l hypothesis.

 ISSN 2476-5880

4.4. Addressing the Second Research Question.

Tee secddd eesearch eee stinn tkkk itt o acctttt tee EFL leaeeess' listenigg cmmeeeeessinn skills as a result of being exposed to purposeful AAL-oriented instruction. To probe the difference between AAL and AOL groups, descriptive and inferential measures were run. To ss s ee eeeciitt ive statistics ttt t tt two goosss l listeii gg cmmeeeeessiss sss s-test are provided below.

Table 9
Descriptive Statistics for the Post-Test of Three Groups

	Groups	N	Mean	Std. Deviation	Std. Error Mean
Post-test	AFL	25	31.00	3.99	.99
	AAL	25	30.50	3.98	.98
	AOL	25	28.00	4.52	1.19

*. The mean difference is significant at the 0.05 level.

The table indicates that AOL has the lowest mean score (28.00). It seems that there is a similarity between AAL and AFL groups, while a difference can be observed between AAL and group with the AOL group.

To compare the mean scores of the above-mentioned three groups, one-way ANOVA was run in Table 10.

Table 10
One-way ANOVA Statistics for the Post-Test of Three Groups

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	421.11	2	233.76	7.11	.00
Within Groups	1341.36	72	27.81		
Total	1762.47	74			

*. The mean difference is significant at the 0.05 level.

As to Table10, it can be concluded that there exists a significant difference among the listening comprehension post-test of three groups (F2, 72= 7.11, p= .003). Thus, it can be inferred that the three groups were significantly different in their listening comprehension, which is an indication of rejecting the null hypothesis. To locate and highlight the difference among the three groups, Tukey HSD Multiple Comparison was run as shown in Table 11.

 ISSN 2476-5880

Table 11
Tukey HSD Multiple Comparison Statistics for the Post-Test of Three Groups

(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
AFL	AAL	0.50	1.01	.12	.98	9.06
	AOL	3.00	1.43	.00	3.79	9.34
AAL	AFL	-0.50	1.01	.12	-9.06	-.98
	AOL	2.50*	0.99	.00	2.84	7.66
AOL	AFL	-3.00	1.43	.00	-3.79	-9.34
	AAL	-2.50	0.99	.00	-2.84	-7.66

*. The mean difference is significant at the 0.05 level.

Table 11 shows that a significant difference can only be observed between AAL and AOL groups (p= 002), while no significant difference is seen between AFL and AAL groups (p= .129). In other words, there was a significant difference between the effects of purposeful AAL-oriented instruction and AOL-oriett ed isstccctinn in eeveliii gg Iaaii an EFL leaeeess' listening comprehension skills.

4.5. Addressing the third Research Question.

Tee tii dd eesearch eee stinn eeee llyigg tee crrr ett eeseacch elll eeed tee EFL leaeeess' listening comprehension skills as a result of being exposed to integrated purposeful assessment-oriented instruction. To probe the difference between Integrated and AOL groups, descriptive and inferential measures were run. To do so, descriptive measures for the listening comprehension post-test of the target groups are provided in Table 12.

Table 12
Descriptive Statistics for the Post-Test of Three Groups

	Groups	N	Mean	Std. Deviation	Std. Error Mean
Post-test	Integrated	25	33.00	3.00	.81
	AAL	25	30.50	3.98	.98
	AOL	25	28.00	4.52	1.19

*. The mean difference is significant at the 0.05 level.

The table indicates that the Integrated group has the highest mean score (M=33.00), while AOL has the lowest one (M=28.00). It can be inferred that there exists a large difference among the three gssssss sisteii cc eemmmmcc.

To compare the mean scores of the above-mentioned three groups for the post-test of listening, one-way ANOVA was run in Table 13.

 ISSN 2476-5880

Table 13
One-way ANOVA Statistics for the Post-Test of Three Groups

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	499.20	2	224.10	7.99	.00
Within Groups	1329.20	72	30.30		
Total	1828.40	74			

*. The mean difference is significant at the 0.05 level.

As to Table13, it can be observed that there exists a significant difference among the listening comprehension post-test of three groups (F2, 72= 7.99, p= .000). Thus, it can be concluded that the three groups acted significantly differently on the post-test of listening comprehension, which indicates that the null hypothesis is rejected. To locate and highlight such differences among the three groups, Tukey HSD Multiple Comparison Statistics was run.

Table 14
Tukey HSD Multiple Comparison Statistics for the Post-Test of Three Groups

(I) Groups	(J) Groups	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Integrated	AAL	2.50	1.11	.00	1.01	4.21
	AOL	5.00	1.1	.00	1.83	6.99
AAL	Integrated	-2.50	1.11	.00	-.1.01	-.4.21
	AOL	2.50	0.99	.00	2.84	7.66
AOL	Integrated	-5.00*	1.1	.00	-1.83	-6.99
	AAL	-2.50	0.99	.00	-2.84	-7.66

*. The mean difference is significant at the 0.05 level.

Table 14 shows the existence of significant differences among Integrated and AAL groups (p= .003), Integrated and AOL groups (p= 000), and AAL and AOL groups (p=.002). In other words, there was a significant difference between the effects of the integrated purposeful assessment scenario (AFL+AAL) and AOL-oriented instruction in developing Iaaii aE EFL leaeeessl listeii gg cmmeeeeensiss s ii lls.

The current investigation adopted a quantitative methodology to meet the objectives of the study. It included the effect of purposeful assessment-oriented instruction on the female pre-intermediate EFL leaeeess' listeii gg cmmrr eeessinn sii lls in aa stt , Iaa. Leaeeess' eesssss es to tee listeii gg cmmeeehessinn test iddicated tee leanness' sigii iicatt impvvvemett in their listening when they were exposed to integrated purposeful assessment. It seems that

tee leaeeess' scccess in tee itt egaat nn AA AFL add AAL can ee a tuucaall e mattea altggggh it might be challenging from another point of view.

Taking the pros and cons of integration, it can be argued that assessment activities are interactively done by both teachers and the learners, which makes the learning environment more cooperative resulting in their learning gains. However, such integration might lead to leaeeess' eccessive ivvll vemett in tee assessmett add leaiii gg ccccess leading to teeir boredmmrr aaaaaaly ii scaaaaging teeir cmmmiii cati... This eemasss tee teaceesss lll e to be cautious not to overflow the use of assessment materials in the classroom and try to control tee .eaeeess' esssss iii lities in iii gg tee assessmett add simll taeesss ly engage in various tasss with teeir eeess. Teaceesss time allccatinn sllll d ee eggieeeeed in eeeer ttt to ttt mcch eeeeee nn tee leanress' sllll eess, wii ch ssss tee iisk ff lacii gg aaa lity leaiii gg add assessment at the same time.

Although integrated assessment seems not to be mentioned in the literature, researchers (Brown, 2008; Chappius, 2009) recommended the integration in more purposeful ways by catt iiii gg tee iii tt that teache''s ii visinn ff leaeeess' eesssss iii lities can m. ee the assessment be at the service of quality learning. The findings of the study are comparable with those of Blaich and Wise (2011) and Wang and Hurley (2012) who supported the emll yymett ff AFL as a sselll isstmmntt in iccreasigg tee leaeeess' eerrrr mance in the classroom and of Archer (2010) and Brown (2008) who strongly concurred that AAL eecggii zed tee leanress' lll es in beigg cnnceeeed witt t eeir eemmmmcc add tyy to sll ve it nn their own, promoting their continuous learning achievements.

It appears that the commonalities between the two make the learning process more communicative for the learners and listening comprehension is achieved under such interactive learning environments when students are engaged in a conscious-based, self-reflective, and autonomous learning environment. As each assessment scenario tries to bold the learning environment by using a variety of assessment tools, it might be rather difficult to have a clear-cut comparison of the two and say which one can be better than the other one.

It might be assumed that those (Blaich & Wise, 2011; Dunn & Mulvenon, 2009) who are in favor of AFL might direct the attention toward the teacher as the assessor of the leaeeess' acii evemett . Teey eelieve taat assessmett is a cmmplex tas,, wii ch suuuld ee nnne by the teacher who is also the provider of both descriptive and evaluative feedback types. On the other hand, AAL supporters (Graziano-King, 2007; Sendzuik, 2010) conversely argue that sharing the responsibilities with the learners can help them construct their identity and develop their metacognitive awareness. Hence, there is no consensus regarding the superiority of one approach over the other, demonstrating that each should be carried out in its aiiiiii i te cnntett acc c ccc eiii gg teachesss all l eanresse eeaagggic add realistic needs.

5. Conclusion

In light of the present study, more areas of inquiry were identified to help multiple stakeholders and consumers. It is necessary to revise and redesign pedagogy to balance the tensions among assessment as, for, and of learning and to use the advantages of each to improve learning and teaching (Mok, 2012). The findings of this research may assist policymakers in emphasizing the significance of the use of different approaches to skills

evaluation. Moreover, it seems that students, teachers, and researchers can also benefit from the outcomes of the present study.

Learners are considered as the first beneficiary of the study findings. Many learners appear to be worried about their listening ability in the process of language learning and are usually concerned with their listening skill as well as their grades in listening exams. Being assessed through a purposeful method of assessment, learners can overcome listening difficulties since they are exposed to practice and interactive learning environment when are productively assessed and consciously involved within the assessment procedure. When learners are aware of their listening skills, they can take necessary action to solve the possible deficiencies in listening as well as strengthening their listening ability and awareness-raising. Since listening comprehension is a demanding task for language learners, purposeful assessment scenarios enable learners to be in charge of their listening progress by self-evaluating their performance while the teacher tries to monitor their learning behaviors GGsss & Simss,,, 2..... . ee leaeeess' ivvll vemett in the assessmett ooccerrr es ttt lll y mtt ivates tee leaeeess' eevelmmmatt ff laggaage sii lls ttt also eelss tee leaeeer to ee awaee of the significance of being assessed during the learning process (Archer, 2010). Besides, assessment scenarios can make the learning process more lasting since keeping track of leaeeess' ggggeess cccsss iiii gg tee temm,wii ch eesll ts in teeir itt eaaalizatinn ff the focused skill (Choi, Nam, & Lee, 2001).

Teachess eeed to vvvviee assessmett eeedaack to mEii mize its ttt ential rrr steeettt s' action in line with the LOA framework. Findings of the present study revealed that teachers might be encouraged to do assessments of AFL- or AAL- oriented instruction to track the leaeeess' ggggeess acciiii gg to tee cccsse ed laggaage sii ll. No matter wii ch laggaage sii ll is concentrated, using purposeful assessment scenarios can provide a neat schedule for teachers to provide feedback for the learners based on their feedback to teachers gathered by their diary writings, interviews, or portfolios (Black & Wiliam, 2006). Findings of the study eemsss taated that tee teaceesss eppettise in vvvviii gg tee mecaaii sm rrr pe٣٣٣٣gg purposeful assessmett sceaariss migtt ee wttt h attett inn sicce wwwwto imll emett ' and wwaat i'' assessmett can be eee ff tee main cass es ff leanness' scccess in iii gg tee listeii gg tasks in the assessment-based learning setting (Buck, 2001).

However, these findings should be generalized with care as the context and sample are not representative of the whole population of English learners in different settings. Moreover, a single, commonly agreed-upon definition of comprehension remains elusive (Cutting & Scarborough, 2006). Different comprehension assessments do not always generalize across items, formats, and subjects due to differing definitions of comprehension.

Therefore, further research is required to explore other variables, such as different learning environments as well as different levels of proficiency, and other language skills. As mentioned by Dobson (2010) and Mok (2012), much more research is required to develop systems of theories and strategies for expanding LOA and to provide evidence of how AOL, AFL, add AAL imvvvve steeettt s' leaiii gg. Meeeove,, AFL, AAL, rr AOL is a staategic ccccess taatvl i ffff hee Geess ffFi tt mmmmteachess' activities. Teachess can be assessmett designers, curriculum developers, and knowledge producers (Zeng, Huang, Yu, & Chen, 2018).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Archer, J. C. (2010). State of the science in health professional education: Effective feedback. *Medical Education, 44*(1), 101-108.

Baaaii, E&& &saiii a, F. .2.1... . lggg walk tl l aggaage assessmett li.eaacy: EFL t. a. eess' reflection on language assessment research and practice. *Reflective Practice*, *20*(6), 745-760.

Bayat, A., Jamshidipour, A., & Hashemi, M. (2017). The beneficial impacts of applying formative assessment on Iranian university students' anxiety reduction and listening efficacy. *Online Submission, 5*(2), 1-11.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5(*1), 7-74.

Black, P., & Wiliam, D. (2006). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. doi:10.1080/0969595980050102

Blaich, C., & Wise, K. (Eds.). (2011). *From gathering to using assessment results: Lessons from the Wabash national study*. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA)

Broadfoot, P., Oldfield, A., Sutherland, R., & Timmis, S. (2014). Seeds of change: The potential of the digital revolution to promote enabling assessment. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing assessment for quality learning* (pp. 373-386). Dordrecht: Springer.

Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education Principles Policy and Practice, 8*(2), 153-169. doi:10.1080/09695940123775

Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students*. New York: Nova Science Publishers

Buck, G. (2001). *Assessing listening.* Cambridge; New York: Cambridge University Press.

Carless, D. (2011). From testing to productive student learning: Implementing formative assessment in Confucian-Heritage Settings. New York: Routledge.

Chappuis, S., & Stiggins, R. J. (2002). Classroom assessment for learning. *Educational leadership: journal of the Department of Supervision and Curriculum Development, N.E.A, 60*(1), 40-43.

Choi, K., Nam, J.-H., & Lee, H. (2001). The effects of formative assessment with detailed feedback on students' science learning achievement and attitudes regarding formative assessment. *Journal of the Korean Association for Research in Science Education, 20*(2), 28-34

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension. *Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured, 10*(3), 277-299. doi:10.1207/s1532799xssr1003_5

Dann, R. (2014). Assessment as learning: Blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice, 21*(2), 149-166. doi:10.1080/0969594x.2014.898128

Dobson, S. (2010). Book review: How assessment supports learning. Learning-oriented assessment in action. *Assessment in Education: Principles, Policy and Practice, 17*(2), 105-112.

Drummond, M. J. (2003). *Assessing children's learning* (2nd ed). London: David Fulton.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research, and Evaluation, 14*(7), 11. doi:10.4324/9780203462041_chapter_1

Earl, L.M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin.

Earl, L. M. (2013). Assessment for learning; Assessment as learning: Changing practices means changing beliefs. *assessment, 80*, 63-71.

Earl, L. M., & Timperley, H. (2014). Challenging conceptions of assessment. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing assessment for quality learning* (pp. 325-336). Dordrecht: Springer.

Farhady, H., Jafarpour, A., & Birjandi, P. (1994). *Testing language skills*. Tehran: SAMT Publications

Fuchs, L.S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199-208

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. New York: Routledge.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference.* Boston, MA: Allyn & Bacon.

Gibbs, G., & Simpson, C. (2005). Conditioss eeee r wii ch assessmett stttttt t steee tt s' learning. *Learning and Teaching in Higher Education, 5*(1), 3-31. Retrieved from http://eprints.glos.ac.uk/3609/

Glazer, N. (2014). Formative plus summative assessment in large undergraduate courses: Why both? *International Journal of Teaching and Learning in Higher Education, 26*(2), 276-286.

Graziano-King, J. (2007). Assessing student writing: The self-revised essay. *Journal of Basic Writing, 26*(2), 75-94.

Harlen, W. (2006). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (pp. 61–80). London: Sage.

Harlen, W. (2012). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (pp. 87–102). London: Sage. https://doi.org/10.4135/9781446250808.n6

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, Mass.: Heinle & Heinle.

Huang, J. (2012). The implementation of portfolio assessment in integrated English course. *English Language and Literature Studies, 2*(4), 15–21

James, M. E. (2006). Assessment, teaching and theories of learning. In J. Gardner (Ed.), *Assessment and learning* (pp. 47-60). London: Sage.

James, M. E. (2013). Educational assessment, evaluation and research: *The selected works of Mary E. James*. London: Routledge.

Lam, R. (2013). Formative use of summative tests: Using test preparation to promote performance and self-regulation. *The Asia-Pacific Education Researcher, 22*(1), 69-78.

Lam, R. (2015). Assessment as learning: Examining a cycle of teaching, learning, and assessment of writing in the portfolio-based classroom. *Studies in Higher Education, 41*(11), 1-18. doi:10.1080/03075079.2014.999317

Lam, R., & Lee, I. (2010). Balancing the dual functions of portfolio assessment. *ELT journal, 64*(1), 54-64. doi:10.1093/elt/ccp024

Lee, I., & Coniam, D. (2013). Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing, 22*(1), 34-50. doi:10.1016/j.jslw.2012.11.003

Lee, I. (2016). Putting students at the center of classroom L2 writing assessment. *Canadian Modern Language Review, 72*(2), 258-280. doi:10.3138/cmlr.2802

L. (2017). *Classroom Writing Assessment and Feedback in L2 School Contexts* (1st ed. 2017 ed.). Springer.

Li, X. (2018). Self-assessment as 'assessment as learning' in translator and interpreter education: Validity and washback. *The Interpreter and Translator Trainer, 12*(1), 1-20. doi:10.1080/1750399X.2017.1418581

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing, 30*(3), 329–344. https://doi.org/10.1177/0265532213480129

Marzano, R. J. (2000). *Transforming classroom grading*. Alexandria, Va: Association for Supervision and Curriculum Development.

McCarthy, M., McCarten, J., & Sandiford, H. (2004). *Touchstone Level 1*. Cambridge: Cambridge University Press.

Medland, E. (2016). Assessment in higher education: Drivers, barriers and directions for change in the UK. *Assessment & Evaluation in Higher Education, 41*(1), 81-96.

Mok, M. M. C. (2012). *Self-directed learning-oriented assessments in the Asia-Pacific*. Dordrecht; London: Springer.

Morrissette, J. (2011). Formative assessment: Revisiting the territory from the point of view of teachers. *McGill Journal of Education, 46*(2), 247-265.

Pham, T. T., & Renshaw, P. (2015). Formative assessment in Confucian heritage culture classrooms: activity theory analysis of tensions, contradictions and hybrid practices. *Assessment & Evaluation in Higher Education, 40*(1):45-59 DOI: 10.1080/02602938.2014.886325

Piaget, J. (1960). *The child's conception of the world*. London: Routledge & Kegan Paul.

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership*, *62*(1), 82–83

Richards, J. C. (2013). *Basic tactics for listening* (3rd ed.). Oxford: Oxford University Press.

Sadeghi, K., & Rahmati, T. (2017). Integrating assessment as, for, and of learning in a large-scale exam preparation course. *Assessing Writing*, *34*, 50-61.

Sendziuk, P. (2010). Sink or swim? Improving student learning through feedback and self-assessment. *International Journal of Teaching and Learning in Higher Education, 22*(3), 320-330.

Stiggins, R. J. (2002). Assessment Crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758-765

Taras, M. (2005). Assessment – Summative and formative – Some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478. doi:10.1111/j.1467-8527.2005.00307.x

Wang, X., & Hurley, S. (2012). Assessment as a scholarly activity? Faculty perceptions of and willingness to engage in student learning assessment. *The Journal of General Education*, *61*(1), 1-15.

Watmani, R., Asadollahfam, H., & Behni, B. (2020). Demystifying language assessment literacy among high school teachers of English as a foreign language in Iran: Implications for teacher education reforms. *International Journal of Language Testing, 10*(2), 129-144.

Wistner, B., Hideki, S., & Mariko, A (nd). An analysis of the Oxford Placement Test and the Michigan English Placement Test as L2 proficiency tests. *Hosei University Repository*

Wei, W. .2.1... . sing smmmative add mmmative assessmett s to evalaate EFL teachess' teaching performance. *Assessment & Evaluation in Higher Education, 40*(4), 611-623.

Xiao, Y., & Yang, M. (2019). Formative assessment and self-regulated learning: How formative assessment supports students' self-regulation in English language learning. *System, 81*, 39-49. doi:10.1016/j.system.2019.01.004

Zarei, N., & Yasami, N. (2016). *The Impact of Formative Assessment and Remedial Teaching on EFL Learners' Listening Comprehension.* Paper presented at the conference proceedings. ICT for language learning.

Zeng, W., Huang, F., Yu, L., & Chen, S. (2018). Towards a learning-oriented assessment to improve students' learning–A critical review of literature. *Educational Assessment, Evaluation and Accountability, 30*(3), 211-250. doi:10.1007/s11092-018-9281-9