PASAA
Volume 54
July - December 2017

Review of Graduate Research on Language Assessment in Turkey between 2011 and 2016

Aysenur Uzun
Ministry of National Education, Turkey
Email: a.uzun15@hotmail.com
Ferit Kilickaya
Department of Foreign Language Education
Mehmet Akif Ersoy University, Turkey
Email: ferit.kilickaya@gmail.com

Abstract

Language assessment is one of the key factors affecting language learning contexts since it not only provides information about the learners and teachers' performance in the classroom but also shapes how languages are learned and practiced depending on how assessment is conducted. Technology, different learner profiles, problems with the existing assessment tools, learners and teachers' needs, and differences in educational settings and policy are some of the leading factors in language assessment. In order to determine the main findings and the current trends on language assessment in Turkey, this study aims to review 25 theses and dissertations that investigate various aspects of English language assessment in Turkey between 2011 and 2016. The total number of graduate studies reviewed on language assessment is 25, and all

of these studies have been included in this review. 21 of the studies are M.A. theses, while 4 studies are the Ph.D. dissertations written on language assessment. These theses and dissertations, which were reviewed in this paper, were obtained from *the Council of Higher Education Thesis Center*, which requires graduate students to upload their studies and allows other researchers to access them. These studies have been reviewed and categorized based on alternative assessment, assessment in tertiary contexts, assessment of young learners, assessment of language skills (speaking and writing), evaluation of EFL exams, and developing assessment tools. The studies have been discussed based on their methodology, together with the main findings obtained, in addition to their strengths and weaknesses. It is believed that examining the recent research on language assessment will be beneficial not only for suggesting ideas to fill the gaps in the literature but also for the researchers who are willing to study language assessment in Turkey.

Keywords: graduate research, language assessment, testing, EFL

Introduction

There are many factors affecting language learning and teaching practices, one of which is language assessment. Thus, this study focused on reviewing 25 theses and dissertations written on English language assessment in Turkey between 2011 and 2016. These theses and dissertations were accessed using the database maintained by *the Council of Higher Education Centre Thesis Center* (2017) (https://tez.yok.gov.tr/UlusalTezMerkezi/); however, one of them could not be reached. Therefore, the copy of dissertation (Toprak, 2015) was requested from the author, and a synopsis of the dissertation was obtained. The theses and dissertations were written at the following universities: Çağ University, Mersin; Atatürk University, Erzurum;

Bilkent University, Ankara; Gazi University, Ankara; Middle East Technical University, Ankara; Pamukkale University, Denizli; Necmettin Erbakan University, Konya; Hacettepe University, Ankara and Marmara University, Istanbul. These dissertations and theses were categorized based on alternative assessment, assessment in tertiary context, assessment of young learners, assessment of language skills (speaking and writing), evaluation of EFL exams, and developing assessment tools. The studies were reviewed and discussed based on these categories. The results of this review might enlighten which aspects and issues of language assessment in the Turkish contexts have received attention in graduate research, indicate the gaps in the methodology in the studies reviewed, and reveal the gaps in the literature. It is hoped that the current review will benefit the researchers in the field of language assessment in Turkey by providing ideas for further research as well as the other researchers around the world to be informed about the status of the studies conducted on language assessment in Turkey.

Alternative Assessment

Under this category, three M. A. theses (Özturan, 2011; Sönmez, 2013; Cirit, 2014), which focus on analyzing strengths and weaknesses of alternative assessment in Turkey, were reviewed. Özturan (2011), in her M.A. thesis, investigated the effects of computer-assisted assessment on students' achievement and their perceptions. The researcher benefited from mixed methodology using a questionnaire, an achievement test, and interviews. The participants of the study were 97 students, who were divided into control and experimental groups. While the experimental group took the exam on computer, the control group took the exam on paper. The results revealed that the experimental group was in favor of using computer-assisted assessment, while the control group was neutral. Moreover, there was a moderate positive relationship between computer literacy and perceived ease of use, which means that the users that are more proficient could use the system without any challenges. However, there was a negative correlation between computer literacy and anxiety. A weak negative correlation between the grade point average and perceived ease of use

was also determined. The students also thought that this system helped them to increase their performances and that they could be more successful. The results showed that the participants in the control group preferred taking the exams on paper. However, nearly half of the participants in the experimental group preferred both types of exams, while the rest of them preferred computer-assisted assessment. The reliability and validity of both types of assessment were similar, indicating that computer-assisted assessment could be used as an alternative to traditional assessment. As a result, this study compared these two assessment types and provided learners' perceptions regarding the use of computer-assisted assessment. However, the instructors' views and the effect of computer-assisted assessment on each language skill were neglected. The researchers, therefore, might focus on these neglected areas in their studies for further research.

Sönmez (2013), in her M.A thesis, investigated the effects of formative assessment on the autonomy of Turkish EFL learners at Karamanoğlu Mehmet Bey University. The participants of this study were 35 preparation class students at the Faculty of Economics and the Faculty of Administrative Sciences. The researcher preferred a mixed methodology through interviews using Autonomy Learner questionnaire and Assessment Preference Scale as pre- and post-tests. In addition, goal setting sheets, self/peer evaluation sheets, giving feedback sheets, and classroom observation checklists were used to collect data. The results indicated that formative assessment affected EFL learners' autonomy positively, that the students were eager to participate, and that the students preferred formative assessment after the implementation of formative assessment process although they had mostly preferred traditional assessment at the beginning of the study. As a conclusion, this study revealed the positive effects of formative assessment; however, not only the students' views but also the teachers' views might have been investigated to triangulate the data. Furthermore, using formative assessment in each language skill and its effects might be the scope of another study.

In another study, Cirit (2014) investigated ELT pre-service teachers' perceptions towards the use of Web 2.0 tools. This study also

investigated ELT teachers' perceptions towards the traditional, alternative, and online assessment methods. The participants were 40 sophomore ELT pre-service teachers at Istanbul University. For data collection, this study used pre- and post-surveys, and reflection papers uploaded to *Edmodo*, and semi-structured interviews. The tools used in this study were *Voki*, *Testmoz*, *Mindomo*, Facebook, *Glogster*, *Prezi*, and *Screencast-O-Matic*. The results revealed that the perceptions of ELT pre-service students towards Web 2.0 tools were positive before the implementation, and there has been an increase in their attitudes. According to the results, the alternative assessment was preferred to online or traditional assessment since the participants felt that the alternative assessment was motivating and provided detailed feedback. Furthermore, the use of alternative assessment improved students' critical thinking skills and increased interaction. The advantages of web 2.0 tools were listed as follows: motivating and effective, practical and enjoyable, timesaving and less stressful, providing feedback and authentic source, and stimulating autonomy. However, the disadvantages of web 2.0 tools were listed as follows: creating technical problems, lack of monitoring, not enhancing learning, time-consuming, challenging, and focus on tools more than the subject. The results revealed that the most favorite feedback types were "whole class evaluation", while the least favorite feedback type was "self-evaluation". Moreover, there were more positive attitudes towards the use of alternative assessment and using technology than the ones towards online assessment. However, negative attitudes towards using traditional assessment methods were also noted down. According to the results, the traditional assessment was easy to use and resulted in an increase in learners' achievement, while online assessment saved time, provided various merits, and it was less stressful. Alternative assessment, on the other hand, focused on the learning process and aimed to assess four language skills. It was found to be more effective based on the responses obtained from the participants. In contrast, online assessment allowed both teachers and students to integrate technology into their lessons, which increased the quality of lessons. Using *Edmodo* for reflections improved the communication between the

students and the teacher. To sum up, this study provided a broad perspective towards assessment methods; however, an experimental study might have been used to compare the effects of specific assessment methods clearly.

Language Assessment in Tertiary Contexts

Of the M.A. theses reviewed, five (Dursun, 2014; Zaimoğlu, 2013; Konkur, 2013; Bayram, 2015; Gönen 2013) studied language assessment in tertiary contexts. In her thesis, Zaimoğlu (2013) aimed to investigate the teachers and students' views on assessment in EFL preparatory school. The study also investigated whether gender, years of teaching experience, education level, and undergraduate institution had an effect on conceptions towards assessment. The researcher adopted a quantitative method using questionnaires given to both the teachers and the students. The questionnaires included *Teacher Conceptions of Assessment Scale* (TCOA), which was used for teachers, while the adopted version of TCOA was used for students as *Conceptions of Assessment Scale* (SCOA). The participants of the study were 400 preparatory school students and 31 teachers at Çağ University in Mersin. The results showed that the teachers mostly preferred "teacher made written tests, student-written works, oral questions, answers, and standardized tests." The results also revealed that gender, educational background, years of experience and institutions that the teachers graduated from had no effect on teachers' perceptions of assessment. However, a significant difference was found only in "school accountability" when the teachers' gender and education level were taken into consideration. Female teachers had lower level perceptions than male teachers had, and the reason might be attributed to the fact that the results of assessments would be used to determine the quality of the schools. In addition to the gender factor, teachers' years of teaching experience had a significant effect on "irrelevance", and it was stated that experienced teachers had a lower mean score for irrelevance than the novice teachers had. In addition to the years of experience, a significant effect was obtained in "improvement" considering the institutions where the teachers

graduated. The results showed that the teachers graduated from the Faculty of Education had more positive attitudes towards "improvement" than those who graduated from other faculties such as Faculty of Arts and Science and Translation. The data collected from students' questionnaires revealed that the type of school that the students graduated from and gender had no effect on conceptions of assessment. As a result, the study showed that both the teachers and the students perceived assessment as a tool for improving learning and teaching. Considering the methodology of the study, using a mixed method rather than only using quantitative method might provide a different perspective on the issue. Conducting interviews with the randomly selected instructors and students might also enlighten the reasons for their choices.

Another study conducted by Dursun (2014) investigated the assessment and evaluation practices used in the Schools of Foreign Languages in Turkey. The researcher also investigated how four skills and subskills were assessed in different universities. She used a quantitative research method by using a questionnaire that included two sections to gather demographic information and the activities used for assessing four skills and subskills. The participants were the students at three private and seven state universities in Turkey. The results indicated that proficiency tests, placement tests, achievement tests, and quizzes were used as assessment and evaluation activities. Moreover, understanding main idea, skimming, information transfer, inferencing, and scanning were the most preferred subskills for assessing listening. The frequently used item types in assessing listening were multiple-choice items, followed by filling blanks, true-false, and fill out forms. To assess reading, understanding main idea, skimming, and referencing were generally used, and the multiple choice item type was the most preferred question format. For speaking assessment, the subskills included a description, having a dialogue on a topic, problem-solving, and given presentations. Having a dialogue on a topic was the frequently used item type, and one student was assessed at a time. For assessing writing, the subskills included description and cause and effect essay writing, while the item type

frequently used in writing appeared to be guided writing. The other skills were vocabulary and language use, and the most preferred item type used in assessing language use was multiple-choice questions, cloze tests, re-writing, and completing dialogues, while the item types such as the cloze test with multiple choice questions and deducing the meaning from context were used in assessing vocabulary. The results also showed that except for speaking, other skills were seen as important by universities since all skills are interrelated to each other and should be assessed with their subskills. Dursun's study investigated the activities used for assessment in universities in detail; however, the mixed methodology can be used rather than using the only quantitative method, and teachers' perspectives might be investigated. Furthermore, some of the exams used for assessment in the universities might be investigated and could be compared with the results of the questionnaire.

Gönen (2013) aimed to investigate the perceptions of the teachers towards classroom-based language assessment and the implementation of assessment in tertiary contexts. The researcher used a mixed methodology that included both questionnaires conducted with 102 teachers and interviews with 5 teachers. The results clearly showed that the participants were well aware of both the aim and planning of classroom-based assessment. The participants stated that their personal characteristics did not affect the students while assessing them. Moreover, the participants stressed that that assessment could not be separated from teaching and that they considered feedback a crucial factor affecting assessment and learning. Therefore, they believed that assessment should not be used as a punishment tool. While the teachers thought that sharing the results with other teachers were not appropriate, they emphasized that positive effect might be seen when the results were shared with the teacher development units. Gönen's study analyzed the results of a classroom-based assessment within teachers' point of views; however, students' views could have been investigated. Increasing the number of interviewees might help the researcher to generalize the findings of the study, and this might be done in further research through replication studies.

In her thesis, Konkur (2013) investigated the students' attitudes and views towards assessment and investigated how to develop an effective learning environment. The participants of this study were 20 preparatory class students at Çağ University, and a qualitative research method was applied. The students' diaries, teachers' notes, observations, and interviews were conducted to collect data. The students in this study took 14 pop quizzes on the contents of the coursebook used, including the four skills in each semester. Furthermore, the students had monthly exams and small grammar quizzes every Monday, and a final exam was given at the end of the academic year. They also received a vocabulary list at the beginning of each week, and they had a vocabulary quiz at the end of the week. In the study, the researcher first asked the students to write sentences, and gradually the students started to write paragraphs. While students were dealing with the tasks, the teacher provided feedback on the students' work. In the first interview conducted before the study, some students did not have enough knowledge on assessment, while some of them thought that the assessment was based on quizzes, exams, and anything done in the class. Even though they attached importance to the quizzes, they did the quizzes to obtain scores to pass the class. In the second interview conducted after the second monthly exam, the students were stressed and panic. They thought that the practices in the class were different from the exam questions even though they were similar. According to the interview results, most of the students tried to memorize the sentences rather than learning, which means that they did not know how to study. Some of them were aware of how to use feedback, while the others were not. In the third interview conducted four weeks before the final exam, there was an increase in the students' awareness, and they started to perceive feedback as an effective way to overcome weaknesses. It was clear that providing feedback increased the students' motivation and helped them to understand how to study. They also had positive attitudes towards the types of the assessment conducted, and most of the students could overcome their stress. When their attitudes towards assessment were positive, they were able to gain self-confidence. As a result, the students could improve their

proficiency levels by studying regularly, benefiting from feedback, increasing motivation, and gaining self-confidence. To sum up, this study provided a general framework of the effective learning environment focusing on the students' attitudes and perceptions. However, the data obtained might have been triangulated by using a quantitative method such as conducting a questionnaire on the students' perceptions of language assessment. As a suggestion for further research, an experimental study including different types of practices might be used to compare and contrast the results obtained.

Bayram (2015), in her M.A. thesis, investigated the pre-service EFL (English as a Foreign Language) teachers' language preferences and underlying factors. This study benefited from a mixed methodology using questionnaires and focus group interviews. After *the Assessment Techniques Awareness* questionnaire, a workshop on assessment techniques was held. The participants were 171 freshmen and 155 seniors in the Department of ELT (English Language Teaching) and ELL (English Language and Literature) at Karadeniz Technical University and Atatürk University; however, only 38 students were selected for focus group interviews. According to the results, the students preferred different language assessment techniques for different reasons based on the factors such as gender, perceived identity, and being a freshman/ sophomore. The difference between the awareness, usage, and preferences was significant, and the participants' preferences were affected by both external and internal factors such as high-stakes tests, teachers, crowded classes, test anxiety, individual differences, and prior knowledge. Their assessment preferences were in favor of mixed language assessment techniques to show their actual performances. It was also clear that they wanted to get detailed feedback, rubrics, and information about language assessment. The results revealed that female students were more aware of assessment techniques, and they were in favor of using performance-based assessment techniques. Female EFL students also preferred to be assessed with language assessment techniques in writing such as creating posters and journals. Therefore, it might be stated that the use and awareness of language assessment techniques might affect their preferences.

Furthermore, extroverted students were more aware of the technique "drama", while introverted students were in favor of using "observation". However, not only the extroverts but also the introverts had similar rates in awareness, and they preferred similar techniques. In addition, the department differences were another issue affecting awareness and preferences. The ELT students were more aware of "structured grid, performance-based assessment, multiple choice questions and rubrics or observations," while the ELL students were more aware of "portfolios and e-portfolios". The ELL lecturers used mostly "translation, journals, portfolios, and peer assessment"; however, ELT students preferred "structured grids and concept maps". It can be stated that the techniques used in the class might affect the students' preferences. Moreover, senior students were more aware of "translation, presentation, oral exams, structured grid, and drama", while freshman students were more aware of portfolios. Presentation, homework, and drama were mostly used for senior students; however, journals, oral exams, portfolios, and peer assessment were used for the freshmen. According to the results, the students preferred being assessed through mixed, alternative, and traditional techniques based on both individual differences and educational requirements. To sum up, this study revealed the effective factors on preferences of students; however, this study might have investigated the lecturers' views to provide a broader perspective regarding the preferences of assessment.

Assessing Young Learners

Assessment of young learners has been investigated in the Ph.D. thesis of Çetin (2011), and M. A. theses of Ayas (2014) and Çiçen (2014). Çetin (2011), in her Ph.D. dissertation, investigated the teachers' practices, students' and teachers' beliefs, and importance of alternative assessment in a private school. The study benefited from a qualitative method by using interviews, observation, and document analysis. The participants of this study were 9 teachers teaching the classes from the first grade to the fifth and 21 students in the third grade. The data collection instruments included two semi-structured interviews conducted with the teachers, focus group interviews with

students, and teachers' observation forms. The results of the study indicated that different types of alternative assessment were used by the teachers, and the most frequently used methods were determined to be observations, portfolios, and self-reflection. It was also found that alternative assessment was affected by planning time, training, classroom environment, language, and cognitive ability. Therefore, it was stressed that motivation and effective factors should be given importance. The teachers believed that students and learning process were affected positively, which led students to become autonomous learners with increased motivation and that alternative assessment enabled getting feedback from the students. However, the challenging issue was the use of alternative assessment sufficiently and effectively. While both school documents and the teachers were aligned with alternative assessment, the results of alternative assessments were not used to realize the objectives. Çetin's study investigated the alternative assessments used with young learners and the perceptions of both the teachers and the students by using several qualitative methods. However, using a mixed methodology and investigating students' perceptions in a detailed way rather than focusing on the teachers' perspectives might provide a better understanding of using alternative assessment.

Another study conducted by Ayas (2014) investigated the perceptions of the teachers and their preferences on language assessment tasks at a state school in Osmaniye. The researcher benefited from a mixed methodology using an assessment questionnaire to determine teachers' conceptions of assessment, which included four subgroups: assessment for improvement, assessment for school accountability, assessment for school accountability, and assessment as irrelevant. The participants of the study were 43 primary school teachers in Osmaniye, and the data were collected via questionnaires. The results of school accountability subgroup indicated that it was a good way to use assessment tools in assessing schools. In addition to this, the teachers also supported the idea of using assessment for grading students and the idea that assessment indicated the quality of school; however, they were opposed to the idea that assessment

indicated the school performance. The teachers also indicated that due to the assessment, they had to teach by using different ways that they did not want to use, and this assessment might affect teaching negatively. According to the teachers, assessment results were convenient and provided feedback to both the teachers and the students. Based on the results, the teachers' choices of assessment tasks were listed as follows: oral questions and answers, standardized tests, written tests, portfolios, self and/or peer assessment, and checklists. The results indicated that teachers' perceptions and their choices on assessment tasks were opposed to each other. Even though they chose traditional tools in assessment, their perceptions of assessment were in favor of using alternative and formative assessment. Ayas's study sheds light on the contrast between the teachers' perceptions and their task choices; however, in addition to the questionnaire, to achieve data triangulation, classroom observation, and interviews might have also been conducted. Furthermore, perceptions and preferences of private and state school teachers might have been compared and contrasted.

Another study that investigated the assessment techniques used by primary school EFL teachers at private school was conducted by Çiçen (2014). The researcher used both qualitative and quantitative methods to collect data, and the participants were 42 EFL teachers working at a private school in Gaziantep. The data collection instrument included a questionnaire with open-ended questions. The results of the study indicated that the teachers mostly preferred 'portfolios' to observe and monitor the teaching process, to increase students' motivation, and to assess student success. In addition to the most preferred method, assessing oral skill appeared to be the most preferred content. The teachers also stated that they preferred authentic assessment, while the students preferred traditional tests. Çiçen's study revealed the teachers' preferences on assessment and students' preferences; however, the researcher did not conduct a questionnaire to investigate students' preferences. Further studies might use both teacher and student questionnaires, and interviews might be conducted to reveal their opinions.

Assessing Speaking

Under the category of assessing speaking, seven M.A. theses (Duran, 2011; Yastıbaş, 2013; Önem, 2015; Bilki, 2011; Köksal, 2013; Karagedik, 2013; Öztekin, 2011), and a Ph.D. Dissertation (Yakışık, 2012) were reviewed. The study, which Duran (2011) conducted with 307 students and 45 instructors at Akdeniz University, investigated the perceptions of both the students and the teachers on the washback effects of the classroom-based speaking test. The researcher preferred a mixed methodology using questionnaires and conducting interviews with six teachers and seven students. The results showed that both teachers and students held positive attitudes towards speaking and stated that speaking was important and could be practiced in class. While the teachers felt that testing speaking was difficult, they also indicated that speaking should be tested through speaking, not writing. They disagreed with the idea that these speaking tests could assess students' speaking skill. Moreover, they were uncertain about the issue whether speaking tests were a reliable tool, and the students were neutral on the idea about the validity and reliability of these tests to test their skills. In addition to the findings on the participants' perceptions, the washback effects of speaking tests were also investigated. The results showed that speaking tests had positive effects on students' speaking skills even though there was no washback effect on teaching, learning, and practices in the class. Thus, the students and teachers emphasized that speaking had a crucial role and speaking tests were beneficial for students as they could produce the language and recognize their weaknesses. Duran's study sheds light on the perceptions towards speaking tests and the washback effects of speaking tests by using a mixed methodology; however, different kinds of techniques in assessing speaking might have been used to compare the effects of these techniques on students' speaking skills and their preferences.

In another study, Yakışık (2012) focused on the effect of the dynamic assessment on ELT learners' speaking skills at Gazi University. The participants of the study were 36 ELT students in the School of Foreign Languages, who were divided into two groups as

experimental and control. The study used a mixed methodology with the student evaluation forms and pre- and post-tests. The students in both groups took pre- and post-tests. However, the students in the experimental group also had L2 enrichment program, transfer assessment session, and student evaluation forms. At the beginning of the study, demographic information was obtained, which was followed by pre-non-dynamic and pre-dynamic assessments. In the assessment procedure, "retelling story test" was used. Then, the implementation session of enrichment program and transfer assessment session were conducted in the experimental group followed by a post-test applied to both groups to see the differences. In the end, student evaluation forms were given to the experimental group to obtain their opinions. The results indicated that both the experimental and the control groups obtained similar results at the beginning of the study; however, the post-tests and the student forms indicated that the experimental group showed a higher improvement and better independent performances compared to the control group. The experimental group was also successful in transferring their abilities to new situations. Moreover, dynamic assessment helped the students to improve their speaking skills, and the experimental group needed less mediation, leading to fewer problems. The experimental group also thought that enrichment program was beneficial for them and that they could improve their speaking skills. In her study, Yakışık investigated the effects of using dynamic assessment and found crucial results; however, teachers' perceptions, in addition to the students' perceptions, might have also been investigated, and dynamic assessment could have been compared with the other types of assessments.

The following study, which was carried out by Yastıbaş (2013), investigated the use and effects of e-portfolio (Lore) in speaking assessment. 17 upper intermediate students in the department of English Language Preparation at Zirve University were the participants of the study. The researcher used a qualitative method by using researcher's diary, students' e-portfolios, cover letters, which were given at the end of the study, interviews, and self-assessment papers, which were conducted at the beginning and at the end of the study. The

results indicated that the students could see their improvements in speaking with the help of self-assessment. Furthermore, due to the group work in the second assignment, students' motivation and creativity were affected positively. As a result, the students improved their self-assessment skills, computer skills, speaking skills and academic skills, and e-portfolios affected the students' attitudes positively in terms of anxiety, self-confidence, and responsibility. However, there were several problems related to the students' computer skills. When these problems are solved, e-portfolios can be used more effectively to assess students speaking skills. As a conclusion, Yastıbaş revealed the issue from the students' perspectives with little attention given to the teachers' perspectives, which might be investigated in a mixed-method research study to reveal more detailed results.

Another study conducted by Önem (2015) investigated instructors' attitudes towards assessing speaking holistically and analytically. The researcher used a questionnaire with multiple choice items and open-ended questions. The participants of this study were 24 language instructors at Erciyes University, School of Foreign Languages. Speaking exams of ten students were recorded, and instructors assessed them holistically and analytically. After the assessment, the questionnaires were given to the instructors, and the data were collected. The results showed that the instructors had more positive attitudes toward holistic assessment. The instructors believed that the merits of holistic assessment were practicality and true reflections of the raters, while the major benefits of the analytic assessment were determined to be providing rich feedback, the reliability of scores, and ease of use. The negative aspects of holistic assessments were determined to be subjectivity, the vagueness of rating process, and requiring training, while the negative aspects of analytic assessment were determined to be time-consuming, having cognitive demand, the gap between the perception of the rater, and calculated scores. The results also showed that there was no significant difference in speaking exam scores assigned using holistic and analytic assessment. Furthermore, the results showed that there was no significant difference between the scores based on the background of

instructors except the years of experience, which indicated that the younger ones assigned higher scores than the older instructors did. Regarding further research, interviews might be used to triangulate the data and students' perceptions of holistic and analytic assessment might be investigated.

While the other studies focused on assessing speaking, Bilki (2011) investigated how effective the cloze tests were in assessing the speaking and writing skills of university EFL learners. The study examined not only the differences in the achievement levels of cloze tests in speaking assessment and writing but also the text selection, deletion methods, and scoring methods to determine the differences in the achievement. The participants were 60 students of the English Language and Literature Department at Celal Bayar University Preparatory School, and the data were collected through six different cloze tests, a speaking exam, and a writing test. The two texts, which were taken from Wall Street Journal and the script of a movie 'The Shining', were turned into cloze tests by using two different types: an article type of cloze and dialogue type of cloze tests, each of which included three methods. Article cloze tests included the deletion of every 13th function word, deletion of every 13th content word and deletion of every 13th word, while dialogue cloze tests included the deletion of every 22nd function word, deletion of every 22nd content word and deletion of every 22nd word. The students were given different deletion types of tests for both text types, and their responses were scored by using two different methods: exact word and acceptable answer scoring methods. The essays in the writing exams were scored by using analytic rubrics by two raters while speaking exams were scored by two raters using the holistic rubric. The results showed that a higher positive correlation existed between the article cloze tests and writing, while a higher positive correlation was found between dialogue cloze tests and speaking. Moreover, the acceptable answer scoring had the highest correlation with both speaking and writing. The results also indicated that the deletion of content words in dialogue cloze tests provided a better assessment in speaking, while the deletion of function words in article cloze tests worked well in writing. Bilki (2011) provided

a different perspective toward assessing speaking and writing by focusing on using cloze tests; however, interviews, students' and teacher's diaries might have been used to triangulate the data. Further studies might also replicate this study in different departments and at different proficiency levels to observe the other effects of cloze tests on assessing speaking and writing.

In another study, Köksal (2013) focused on the rater reliability in oral interview assessments. The researcher investigated the effects of raters' prior knowledge of students' proficiency levels on scoring. This study was a quasi-experimental study and the researcher preferred mixed methodology using pre- and post-tests, and think aloud protocol sessions. The participants of the study were 15 EFL instructors. The study used six videos from the proficiency exam as the data collection materials, and each video included the interviews of two students. First, the instructors assigned scores to four students in two extra recordings in the morning session. Then, they used an analytic rubric to assign scores to students both in the pre- and post-tests and the instructors verbalized their thoughts while assigning scores to students at three different levels. In order to investigate the effects of raters' prior knowledge of students', the instructors were informed about students' level in the post-test, while they were not informed in the pre-test. The results revealed that more than half of the instructors changed the scores that they had assigned in the pre-test. The study indicated that the reason might be that they might guess the performance of the students based on their level. According to the results, the instructors tended to give higher scores in the post-test to the higher-level students, while they tended to give lower scores in the post-test to the lower-level students. To sum up, the leniency or severity degree of raters was affected by the students' proficiency levels. Based on the neglected areas of this study, the interviews and questionnaires might have been used to triangulate the data in addition to the think-aloud protocol analysis and the use of the pre- and post-tests. Moreover, further research should also make sure that the student pairs in the oral exam should have same proficiency level as the interaction of two students at different levels might affect the raters' scoring.

Another study conducted by Karagedik (2013) focused on the instructors' needs in teaching speaking. The researcher investigated the objectives, content, teaching/learning procedure, and assessment of the in-service training program "Teaching Speaking Skills for English Instructors". This study was an action research study conducted with 11 instructors at Ankara University School of Foreign Languages. The study benefited from questionnaires, interviews, achievement tests, and observations to collect the data. Furthermore, the researcher designed an in-service training program based on the findings and assessed it after the training program was completed. Based on the table of specifications created by using the objectives and contents of the program, the achievement tests were used as the pre- and post-tests. The results indicated that the instructors needed some guidance regarding the grading process and to be informed about the students' proficiency level. They needed guidance and training on how to interact with the students and determine materials based on the students' interests. The results also revealed that the instructors needed to be guided how to take part in activities as a participant or an observer. They needed to be informed not only how to group the students but also how to encourage them in their presentations. Moreover, the instructors also expressed that they needed some guidance about giving feedback, and self- and peer assessment. This study provided a detailed analysis of the instructors' needs in teaching speaking skill. However, not only the instructors' views but also the student's views regarding the outcomes of the training program might have been investigated to provide more valuable insights.

In her M.A. thesis, Öztekin (2011) compared computer-assisted and face-to-face speaking assessment (FTFsa) based on the participants' performance, perceptions, anxiety, and attitudes towards the use of computers. In addition, this study investigated both the advantages and disadvantages of computer-assisted speaking skill (CASA) and FTFsa on speaking assessment. The researcher preferred mixed methodology using CASA, FTFsa, a speaking anxiety questionnaire, a computer familiarity questionnaire, and a questionnaire on perceptions towards CASA and FTFsa. This study was

conducted at Uludağ University, and the participants were 4 instructors and 66 students at the School of Foreign Languages. There were two groups of students at intermediate and pre-intermediate levels. Furthermore, these two groups were divided into two once again to apply FTFsa and CASA. In the first speaking test, Group-1 pre-intermediate students took FTFsa, while Group-2 pre-intermediate students took CASA as Test-1. Group-1 intermediate students took CASA, while Group-2 intermediate students took FTFsa. In the second test, Group-1 pre-intermediate students took CASA, while Group-2 pre-intermediate students took FTFsa. Group-1 intermediate students took FTFsa, while Group-2 intermediate students took CASA as Test-2. There was a one-month interval between the first and the second speaking test. According to the results, students' scores were not affected by the test types. The results also revealed that pre-intermediate students performed better in FTFsa, while intermediate level students scored higher in CASA.

However, this difference was not significant, and there was no correlation between the scores on two test types at both levels. The results indicated that test type, level, or group alone did not affect the students' scores. However, the second test affected the scores positively, and this was the result of "practice effect". This practice effect was only observed at the pre-intermediate level. The results also revealed that the pre-intermediate level students preferred FTFsa, and had more positive attitudes toward it, while the intermediate level students' preferences were similar to each other, and they preferred CASA. This might be due to intermediate students' lower anxiety and the increased self-confidence. However, both groups felt more anxious in CASA, and this might be attributed to the technical problems, students' unfamiliarity with this assessment form, lack of opportunity to ask for clarification and repetition. The results also showed that pre-intermediate level students who felt more anxious might have less positive attitudes toward CASA, but the perceptions toward FTFsa and anxiety were not related to each other. Even though there was a negative correlation between pre-intermediate level students' computer attitudes and CASA perceptions, there was a positive correlation

between computer attitudes and CASA scores. Intermediate level students who felt more anxious were found to have less positive attitudes toward both CASA and FTFsa. There was a negative correlation between the scores of intermediate level students in CASA, FTFsa, and anxiety. Consequently, anxiety was determined to be related to the perceptions toward FTFsa at the intermediate level, but not at the pre-intermediate level. As a result, this study provided a broad perspective toward using CASA and FTFsa; however, the researcher might have used learners' logs or interviews to triangulate the data to provide a different perspective.

Assessing Writing

In addition to assessing speaking, assessing another productive skill, writing was investigated in the Ph. D. dissertation of Han (2013), and M.A. theses of Doğan (2013) and Banlı (2014).Han (2013) investigated the effect of using different methods (holistic and analytic) and rater training on the reliability and validity on EFL students' writing skill. The participants were 36 students and 19 raters at the Department of English Language and Literature at a state university. This study was carried out as an experimental and natural context study that used non-random convenience sampling strategy for selecting participants. This study also benefited from a quantitative method based on G-theory and a qualitative method based on the interviews with raters. 10 raters were in the experimental context, while 9 raters were in the control context. The students wrote essays on two topics, and there were 72 essays in total, which were first analyzed holistically and then analytically by ten trained raters in the experimental group. In addition, the same 72 essays were rated both holistically and analytically by nine raters. These essays were used to examine the effect of rater training on reliability and variability of the scores assigned. The follow-up interviews included four raters in the experimental study and four raters in the natural context. The data collected from both the experimental and the control groups indicated that there was no significant difference between the analytic and holistic scores. The results also showed that the holistic scoring method

was as reliable as the analytic one; however, rater training had an effect on reliability and variability of EFL writing scores. Moreover, the interview results indicated that it was challenging for the raters to select the scoring methods, each of which had weaknesses and strengths. It was determined that scoring rubrics had an effect on raters' scoring. Thus, the raters used analytic rubrics more frequently rather than holistic rubrics. This study focused on the instructors' perspectives toward scoring methods; however, the gender factor in rating, students' perceptions, and preferences might have also been investigated by using questionnaires or conducting interviews.

Another study conducted by Doğan (2013) focused on the "Automated Essay Scoring System" and investigated whether this system was effective and useful in assessing writing. It was a quantitative study, and 50 students at the School of Foreign Languages at Zirve University were the participants of the study. The data were collected from students' essays in the final exam, and these essays were graded by three human raters in four days, while "Criterion", which was a holistic automated essay scoring system, graded them just in one hour. Moreover, the data gathered both from both the human raters and the e-rater were compared to determine the effectiveness of the automated essay scoring system. The results of the study revealed that the automated essay scoring system helped the teachers to give feedback to students and monitor student progress. In addition, less time was used by the system when compared with the human raters. The results also showed that this system was as valid and reliable as the human raters were. To sum up, Doğan investigated the effectiveness of the automated scoring system, suggesting it to be used in writing assessment due to its merits such as saving time, being valid and reliable, and providing feedback. However, the teachers' perceptions toward the system might have been investigated through interviews. Moreover, an analytic method system as an e-rater might have been used to compare it with the holistic e-rater.

In the next qualitative case study, Banlı (2014) investigated the effect of self-assessment on students' writing skills and their awareness. The participants of the study were elementary level 22

Automotive Engineering freshmen at Mersin University. The instruments of the study were the self-reflection checklist given at the end of each writing session, student and teacher journals, and open-ended questionnaires. The results indicated that more than half of the students assessed their performances successfully, and self-assessment helped them to become aware of their strengths and weaknesses. In addition, self- assessment helped them to set goals and provide feedback, which was crucial to improving their writing skill. Besides the effects of self-awareness, task awareness was also investigated, and the results revealed that feedback on teaching and learning was obtained with the help of checklists and journals, thus enabling teachers to monitor the improvements in the students' performances. As a result, Banlı showed the importance of self- assessment in the writing by focusing on the students' perspective; however, an experimental study might be conducted to compare the weaknesses and strengths of self-assessment. Apart from this, a quantitative method might support the study by revealing different aspects of self- assessment on writing such as teachers and students' attitudes towards self-assessment, the washback effects of self-assessment, and its effects on learners' autonomy.

Evaluation of EFL exams

The only study conducted on EFL exams was carried out by Gürsoy M.A thesis published in 2013. In her thesis, Gürsoy (2013) investigated the reliability and validity of the English proficiency exam at Çağ University, which included vocabulary, listening, writing, reading, and grammar. The researcher used both qualitative and quantitative methods to collect data, and the participants of the study were 133 preparatory school students. The students' exam results were analyzed using descriptive statistics and were subject to content analysis. Moreover, the test content and components were analyzed by three instructors at Çağ University using Bachman's model for language ability. The results of the study showed that the content of English proficiency exam matched the principles of communicative language teaching. The results also indicated that the high reliability

was observed in grammar and vocabulary, while the highest reliability was found in the listening and reading. The study also showed that listening and reading had good reliability coefficient, and they could be useful in assessing listening and reading. From the point of item discrimination and item difficulty, the discrimination level of items in grammar and vocabulary was determined to be low, and the items in grammar were found to be difficult, requiring the revision of these items. The items in dialogue completion questions were found to be difficult, and this was attributed to the students' lack of sociolinguistic competence. Thus, oral exams were considered more beneficial than multiple choice formats in assessing sociolinguistic competence. To sum up, Gürsoy analyzed one of the English proficiency exams conducted at the tertiary level by investigating the validity and reliability of the exams. However, authenticity and washback effects of the exam might have also been investigated.

Developing Assessment Tools

Under the category of developing assessment tools, two studies were reviewed: Toprak's Ph.D. dissertation published in 2015 and Başer's M.A. thesis in 2015. Başer (2015) developed and evaluated a *Technological Pedagogical Content Knowledge* (TPACK) assessment tool for pre-service English language teachers. The researcher used a mixed method to develop and evaluate this tool, and 88 pre-service teachers were the participants of this study. TPACK- EFL survey was developed based on the interviews conducted with the experts, analyzing national and international standards, and existing TPACK surveys. Moreover, interviews were conducted with 7 instructors and an EFL pre-service teacher to ensure the content validity of TPACK- EFL survey. After the construct validity of TPACK-EFL survey was evaluated, the interviews were conducted. The participants of the study were asked to complete the survey, which was followed by the interviews conducted with 12 pre-service teachers. The results indicated that most of the items were aligned; however, 3 misaligned TCK and TPK items were determined to be improved. This study also showed that there was a high level of convergence within TPACK construct. Further research might benefit

from learner/teachers logs and think-aloud protocols to provide more detailed information on the stages of development and participants' perceptions.

Another study conducted by Toprak (2015) aimed to create a cognitive diagnostic test, and the test data were analyzed using a *Diagnostic Classification Model* (DCM) and Log-linear *Cognitive Diagnosis Model* (LCDM). The participants of the study were 1058 ELT students. The researcher used a mixed method approach to the study and used Cognitive *Diagnosis Assessment of Second Language Reading Comprehension* (CDSLRC) test as a tool and the product of the study. This test included 5 reading passages, and these passages consisted of 27 multiple choice questions. The test was piloted, and a self-report questionnaire was given to the participants. Moreover, think-aloud protocols with 15 students from upper and lower performing groups on CDSLRC test were carried out, and follow up interviews were conducted. The LCDM framework was used in this study, and the estimations were done using *M Plus*. In addition, syntax analyses were conducted by using *SAS 9.4*. The results indicated that this test was appropriate for its purpose, and the quality of items was determined to be satisfactory. Besides the quality of items, the test discriminated the skilled and less skilled readers. It was also determined that the content validity and the quality of the test were good in terms of the diagnostic feedback provided. As a result, the researcher provided an appropriate test that could be used as a diagnostic test to determine learners' weaknesses and strengths and triangulated the data using both qualitative and quantitative methodology.

Conclusion

The current review revealed that the theses and dissertations written between 2011 and 2016 focused on the assessment in tertiary contexts, assessment of young learners, assessing speaking and writing, developing assessment tools, and evaluation of EFL exams. The studies under these categories revealed the trend of language assessment in Turkey and the underlying factors influencing this trend between 2011 and 2016. The changes in the educational policy of

Turkey such as the frequent changes in the curriculum and in the high-stakes exams are definitely the main factors affecting the language assessment, and these changes are going to start new and different trends in language assessment. Based on the studies reviewed, it is apparent that several other skills and language components such as listening, speaking, reading, grammar, and vocabulary have not received much attention and been little concern to graduate research. This might be attributed to the frequent changes introduced to the curricula as well as the high-stakes exams, which do not assess several of these skills. Developing and improving the current and/or new assessment tools in assessing these skills might be the focus of further research. Assessing speaking and listening, especially in the classes of young learners, surely must be given importance as it is believed that the early exposure to the target language and the input provided during the early years play an important role in the future success and motivation of the learners. The students in Turkey take multiple-choice format exams based on their reading and vocabulary knowledge, and further research can focus on assessing reading and vocabulary through alternative and other types of assessment such as various forms of formative and summative assessment. Based on the results of the studies reviewed, it has been determined that the most striking and problematic issue in language assessment in Turkey appears to be related to assessing speaking and been noticed that assessing speaking appears to be given importance at the territory level, which is the result of educational and language learning policy in Turkey. In school settings, the communicative functions of language and productive skills do not receive much attention, if not any. Therefore, further research can focus on the ignored language skills, together with the language components. The researchers might also consider the suggestions pointed in the current review of the studies. In addition, there are several English language exams in Turkey such as *ÖABT* (Teaching Field Knowledge Exam), *YDS* (Foreign Language Exam), e-YDS (Electronic Foreign Language Exam) and *YÖKDİL* (Council of Higher Education Foreign Language Exam). Therefore, the reliability, validity, and washback effects of these exams might be investigated in further

research. It is hoped that this review will be beneficial for the readers, as well as the researchers, in the field of language assessment to review the current research practices and notice the gaps in the literature, which might give some ideas for further research.

Acknowledgement

This article is the extended version of the paper presented at the 3rd International Language, Culture & Literature Symposium at Akdeniz University, Turkey, on June 15, 2017.

The Authors

Aysenur Uzun is a graduate student at the Department of Foreign Language Education, Mehmet Akif Ersoy University, Turkey. She received her B.A. in English Language Teaching at Mehmet Akif Ersoy University in 2015. Her main area of interests includes language assessment, and teaching English to young learners. She can be reached at a.uzun15@hotmail.com.

Dr. Ferit Kilickaya is currently working at the Department of Foreign Language Education, Mehmet Akif Ersoy University, Turkey. He received his M.A. and Ph.D. degrees in English Language Teaching at Middle East Technical University. His main area of interests includes computer-assisted language learning (CALL), teacher education and technology, language teaching methodology, second language education, language testing, authoring tools, and culture and language teaching. He can be reached at ferit.kilickaya@gmail.com.

References

Ayas, N. (2014). *An investigation of teachers' conception and practices concerning assessment in English for young learners' classroom.* Unpublished Master's thesis, Çağ University, Mersin, Turkey.

Banlı, S. (2014). *The role of self-assessment practices in the improvement of freshman students' writing performance and awareness* Unpublished Master's thesis, Çağ University, Mersin, Turkey.

Başer, D. (2015). *Development and evaluation of a technological pedagogical content knowledge (TPACK) assessment tool for preservice teachers learning to teach English as a foreign language.* Unpublished Master's thesis, Middle East Technical University, Ankara, Turkey.

Bayram, E. (2015). *An investigation of language assessment preferences of pre-service EFL teachers and underlying factors.* Unpublished Master's thesis, Karadeniz Technical University, Trabzon, Turkey.

Bilki, U. (2011). *The effectiveness of cloze tests in assessing the speaking/writing skills of university EFL learners.* Unpublished Master's thesis, Bilkent University, Ankara, Turkey.

Cirit, N. C. (2014). *Perceptions of ELT pre-service teachers toward alternative assessment via web 2.0 tools: A case study at a Turkish state university.* Unpublished Master's thesis, Middle East Technical University, Ankara, Turkey.

Council of Higher Education Thesis Center. (2017). https://tez.yok.gov.tr/UlusalTezMerkezi/

Çetin, L. M. B. (2011). *An investigation into the implementation of alternative assessment in the young learners' classroom.* Unpublished Doctoral thesis, Middle East Technical University, Ankara, Turkey.

Çiçen, M. (2014). *The assessment methods used by EFL teachers at the primary level in private schools.* Unpublished Master's thesis, Çağ University, Mersin, Turkey.

Doğan, A. (2013). *Automated essay scoring system: A reliability study* Unpublished Master's thesis, Çağ University, Mersin, Turkey.

Duran, Ö. (2011). *Teachers' and students' perceptions about classroom-based speaking tests and their washback*. Unpublished Master's thesis, Bilkent University, Ankara, Turkey.

Dursun, G. (2014). *An analysis of assessment and evaluation activities in the school of foreign languages in Turkey.* Unpublished Master's thesis, Pamukkale University, Denizli, Turkey.

Gönen, K. (2013). *Teachers' perceptions of classroom based language assessment in tertiary level English language programs in Turkey*. Unpublished Master's thesis, Çağ University, Mersin, Turkey.

Gürsoy, S. (2013). *The English proficiency exam in EFL context: A validation study*. Unpublished Master's thesis, Çağ University, Mersin, Turkey.

Han, T. (2013*). The impact of rating methods and rater training on the variability and reliability of EFL students' classroom based writing assessments in Turkish universities: An investigation of problems and solutions*. Unpublished Doctoral thesis, Atatürk University, Erzurum, Turkey.

Karagedik, E. (2013). An action research regarding the training programme "Teaching Speaking Skills" for English instructors. Unpublished Master's thesis, Hacettepe University, Ankara, Turkey.

Konkur, B. G. (2013). *Developing an effective learning environment in an EFL classroom.* Unpublished Master's thesis, Çağ University, Mersin, Turkey.

Köksal, T. F. (2013). *The effect of raters' prior knowledge of students' proficiency levels on their assessment during oral interviews.* Unpublished Master's thesis, Bilkent University, Ankara, Turkey.

Önem, E. E. (2015*). Instructors' attitudes towards assessing speaking holistically and analytically*. Unpublished Master's thesis, İhsan Doğramacı Bilkent University, Ankara, Turkey.

Öztekin, E. (2011). *Comparison of computer assisted and face to face speaking assessment: Performance, perceptions, anxiety, and computer attitudes*. Unpublished Master's thesis, Bilkent University, Ankara, Turkey.

Özturan, T. (2016). *The impact of computer assisted assessment on the exam success and attitudes of prospective English teachers.* Unpublished Master's thesis, Hacettepe University, Ankara, Turkey.

Sönmez, T. (2013). *The effects of formative assessment on learner autonomy of Turkish adult EFL learners.* Unpublished Master's thesis, Necmettin Erbakan University, Konya, Turkey.

Yakışık, B. Y. (2012). *Dynamic assessment of ELT students' speaking skills*. Unpublished Doctoral thesis, Gazi University, Ankara, Turkey.

Yastıbaş, A. E. (2013). *The application of e-portfolio in speaking assessment and its contributions to students' attitudes towards speaking*. Unpublished Master's thesis, Çağ University, Mersin, Turkey.

Toprak, T. E. (2015). *Cognitive diagnostic assessment of second language reading comprehension: Application of the Log-Linear cognitive diagnosis modeling to language testing.* Unpublished doctoral dissertation, Gazi University, Ankara, Turkey.

Zaimoğlu, S. (2013). *Teachers' and students' conceptions of assessment in a university EFL preparatory school context*. Unpublished Master's thesis, Çağ University, Mersin, Turkey.